The crystal structure of neuraminidase from Micromonospora viridifaciens

Gaskell, Andrew

Award date:
1995

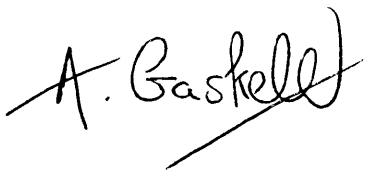
Awarding institution:
University of Bath

THE CRYSTAL STRUCTURE OF NEURAMINIDASE FROM *Micromonospora viridifaciens*

Submitted by Andrew Gaskell for the degree of Ph.D. of the

University of Bath, 1995

COPYRIGHT



Andrew Gaskell

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

The non-pathogenic actinomycete *Micromonospora viridifaciens* secretes a 41kDa neuraminidase when colominic acid is used as an inducer and a 68kDa neuraminidase when milk casein is used. The X-ray crystal structure of the 41kDa form of the enzyme was solved by the technique of multiple isomorphous replacement and the structure of the 68kDa enzyme was solved using molecular replacement with weighted difference Fourier maps.

The 41kDa form of the enzyme has been refined to a resolution of 1.8Å with a crystallographic R-factor of 17.3%. The enzyme exhibits the canonical β-propeller fold found in influenza virus neuraminidase and the bacterial neuraminidases from *Salmonella typhimurium* and *Vibrio cholerae*. In addition the structure of the 41kDa enzyme complexed with the inhibitor 2-deoxy-2,3-dehydro-N-acetylneuraminic acid has been solved and refined to a resolution of 2.0Å with a crystallographic R-factor of 15.5%. The enzyme retains the same conserved active site residues found in the other neuraminidases, although there are subtle differences. In spite of a low sequence identity between the 41kDa form of the neuraminidase from *M.viridifaciens* and other neuraminidases, the overall fold of the enzyme is remarkably similar.

The 68kDa enzyme has been refined to a resolution of 2.5Å with a crystallographic R-factor of 21.4%. In the 68kDa form of the enzyme, the catalytic domain is followed by a "linker" domain which has the immunoglobulin fold and a third domain which has high homology to the galactose binding domain of galactose oxidase from *Dactylium dendroides*. The bacterium may have acquired these immunoglobulin and galactose binding modules from eukaryotes by horizontal gene transfer. The

galactose binding site of the 68kDa enzyme was established by collecting data from the complex and the structure of the 68kDa enzyme with bound galactose has been refined to a resolution of 2.5Å with a crystallographic R-factor of 22.8%. Putative functions for the additional domains of the 68kDa enzyme are suggested and discussed.

Acknowledgements

Firstly I would like to thank my supervisor Garry. For his patience and continued support and enthusiasm throughout my three years at Bath I will be forever indebted (grovel). Susan was of immense help during my time at Bath and always found time to come to my rescue. Peter's immense knowledge of the computing side of things was an invaluable aid to me. Jamie, who has now left the lab, also added to the comedy atmosphere. Pert, UBSA companion and fellow colleague, was of great moral support to me during my time here. There are too many others, to whom I owe thanks, to mention all of them here. However I feel that I must acknowledge Ms. Sometimes, Jodphur Jane, Stella and Little Chick for hours of laughs. Roz, Maria and Ashley, for Italian fun. Shrove boy and Lizzy for punting fun. My housemates at this time, Chris and Simon, for golfing adventures as well as Kev, Joanne, Bob and mad woman.

Abbreviations

C.perfringens	*Clostridium perfringens*
C.septicum	*Clostridium septicum*
D.dendroides	*Dactylium dendroides*
DANA	2-deoxy-2,3-dehydro-N-acetylneuraminic acid
E.coli	*Esherchia coli*
Fn3	Fibronectin type 3 module
FFT	Fast Fourier transform
FPLC	Fast protein liquid chromatography
Gal	Galactose
GCG	Genetics computer group (Wisconsin)
Glc	Glucose
GPI	glycosyl phosphatidylinositol
kD	Kilo Dalton
K_i	Inhibition constant
MHC	Major histocompatibility complex
MIR	Multiple isomorphous replacement
M.viridifaciens	*Micromonospora viridifaciens*
M	Molar
MAD	Multiple wavelength anomalous diffraction
NANA	N-acetyl neuraminic acid
Neu5Ac2en	2-deoxy-2,3-dehydro-N-acetylneuraminic acid
PAGE	Polyacrylamide gel electrophoresis
PEG	Polyethylene glycol

RMS	Root mean square
S.typhimurium	*Salmonella typhimurium*
SDS	Sodium dodecyl sulphate
S_N1	Nucleophilic substitution type one reaction
SWISSPROT	Swissprotein sequence database
TRIS (TRIZMA base)	Tris[hydroxymethyl]aminoethane
T.brucei	*Trypanosoma brucei*
T.cruzi	*Trypanosoma cruzi*
V.cholerae	*Vibrio cholerae*
V_m	Volume in the crystal occupied by one Dalton
NBS	N-bromosuccinimide
PCMB	p-chloromercuribenzoate
EDTA	ethylenediaminetetraacetic acid

Symbols

a, b, c, α, β, γ, Real space unit cell axis lengths and interaxial angles

B, Isotropic temperature factor

$\mathbf{F}_{obs}$, F_{obs}, Observed structure factor and its amplitude

x, *y*, *z* fractional coordinates

ρ (x, y, z) electron density at a point x, y, z in the unit cell

$\mathbf{F}_{calc}$, F_{calc}, Calculated structure factor and its amplitude

h, k, l, Miller indices

f_j, Atomic scattering factor

R, Crystallographic residual

V, Unit cell volume

w, weighting parameter

x, y, z, Cartesian coordinate axes

I, Intensity

α, phase angle

α, β, γ, Eulerian angles

λ, wavelength

σ, standard deviation

FOM, figure of merit

Amino acids: single and triple letter codes

Alanine	ALA	A	Lysine	LYS	K
Arginine	ARG	R	Leucine	LEU	L
Asparagine	ASN	N	Methionine	MET	M
Aspartic acid	ASP	D	Phenylalanine	PHE	F
Cysteine	CYS	C	Proline	PRO	P
Glutamine	GLN	Q	Serine	SER	S
Glutamic acid	GLU	E	Threonine	THR	T
Glycine	GLY	G	Valine	VAL	V
Histidine	HIS	H	Tyrosine	TYR	Y
Isoleucine	ILE	I	Tryptophan	TRP	W

Programs used

RESEC - Resections a given electron density map.

EXTEND - Extends the asymmetric unit to cover any grid volume.

XDS - program by Kabsch to process data collected on the area detector.

DENZO - program by Otwinowski to process data collected on film and the image plate.

XSCALE - program by Kabsch to merge data sets.

X2L* - converts hkl files to LCF format.

LCF2MTZ* - converts LCF format to MTZ format.

MTZUTILS* - combines two reflection files and can be used to edit columns, labels or titles.

MTZDUMP* - lists the header and reflections of an MTZ formatted reflection file.

MTZTONA4* - converts MTZ files to NA4 ASCII format.

CAD* - combines and sorts a number of reflection files. It was used to convert data from one area of reciprocal space to another.

LOCAL - Local scaling of two data sets.

FFT* - performs the crystallographic fast Fourier transform.

PLUTO* - used to plot sections of Patterson maps.

VECSUM* - automated search for Patterson peaks.

VECREF* - Patterson space refinement of the heavy atom sites.

MLPHARE* - refines heavy atom paramters and calculates phases.

FHSCAL* - scales native to derivative data.

SIGMAA* - phase combination of isomorphous and calculated phases.

PEAKMAX* - searches for peaks in an electron density map.

SFALL* - calculates structure factors from an electron density map or atomic coordinates using the inverse FFT.

RSTATS* - scales F_o to F_c using least squares.

MOLSCRIPT - a drawing package.

RASTER 3-D - renders the output from MOLSCRIPT.

DSSP - determines the secondary structure of the protein from a set of atomic coordinates.

CONTACT* - lists the contacts between specified atoms in a structure.

GCG - University of Wisconsin Genetics Computer Group: a sequence analysis software package.

ALSCRIPT - program to improve the presentation of multiple sequence alignments, by Geoff Barton.

COMPOSER - molecular modelling software from Birkbeck. MNYFIT from COMPOSER was used here to perform structural superpositions.

WANG - algorithm by the aforementioned to perform solvent flattening of electron density maps.

SQUASH - density modification package. Can perform solvent flattening, histogram matching and uses Sayer's equation to phase extend.

* programs from the CCP4 (Collaborative Computing Project Number 4) suite.

Contents

Chapter 1

INTRODUCTION

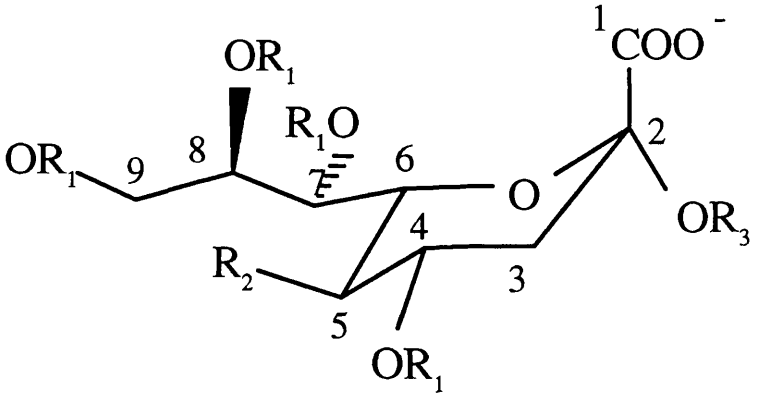
Under investigation is the enzyme neuraminidase (or sialidase), from the non-pathogenic actinomycete *Micromonospora viridifaciens*. The terms sialidase and neuraminidase are generally interchangeable, since the enzyme hydrolyses the removal of sialic or N-acetyl neuraminic acid. A detailed account of this group of enzymes is given, emphasizing the pivotal role in nature which neuraminidases play, followed by some background to the organism from which the neuraminidase in question originates and the aims of the project.

A brief history

The first neuraminidases to be discovered were from influenza virus and *Vibrio cholerae* in 1940. It was found that allantoic fluid from chick embryos infected with influenza virus agglutinated red blood cells in the cold. However, when the temperature of the cells was raised to 37 °C, the cells dispersed and could not be re-agglutinated with further virus. This suggested that the surface of the red blood cells contained a substrate for an enzyme in the virus, which was also the receptor for the virus itself. It was then also discovered that these virus receptors could be removed by an enzyme secreted by *V.cholerae* and the molecule split from

the virus receptors was identified as sialic acid (Hirst, 1941; Gottschalk & Lind, 1949). Thus, neuraminidase catalyses the cleavage of terminal sialic acid residues from glycoconjugates and oligosaccharides. The 9-carbon backbone common to all sialic acids is shown in fig. 1.1, with possible substitutions.

Fig. 1.1 Diversity in the sialic acids. Natural substitutions observed to date at C_4 - C_9 are indicated. Various types of glycosidic linkage at R_3, the generation of lactones at C_1, dehydro forms by eliminating R_3 and anhydro forms generate additional diversity.



R_1 = H, Acetyl (4,7,8,9), Lactyl (9), Methyl (8), Sulfate (8), Phosphate (9), Anhydro (4,8 or 2,7), Sialic acid (8,9), Fucose (4), Glucose (8), or Galactose (4)

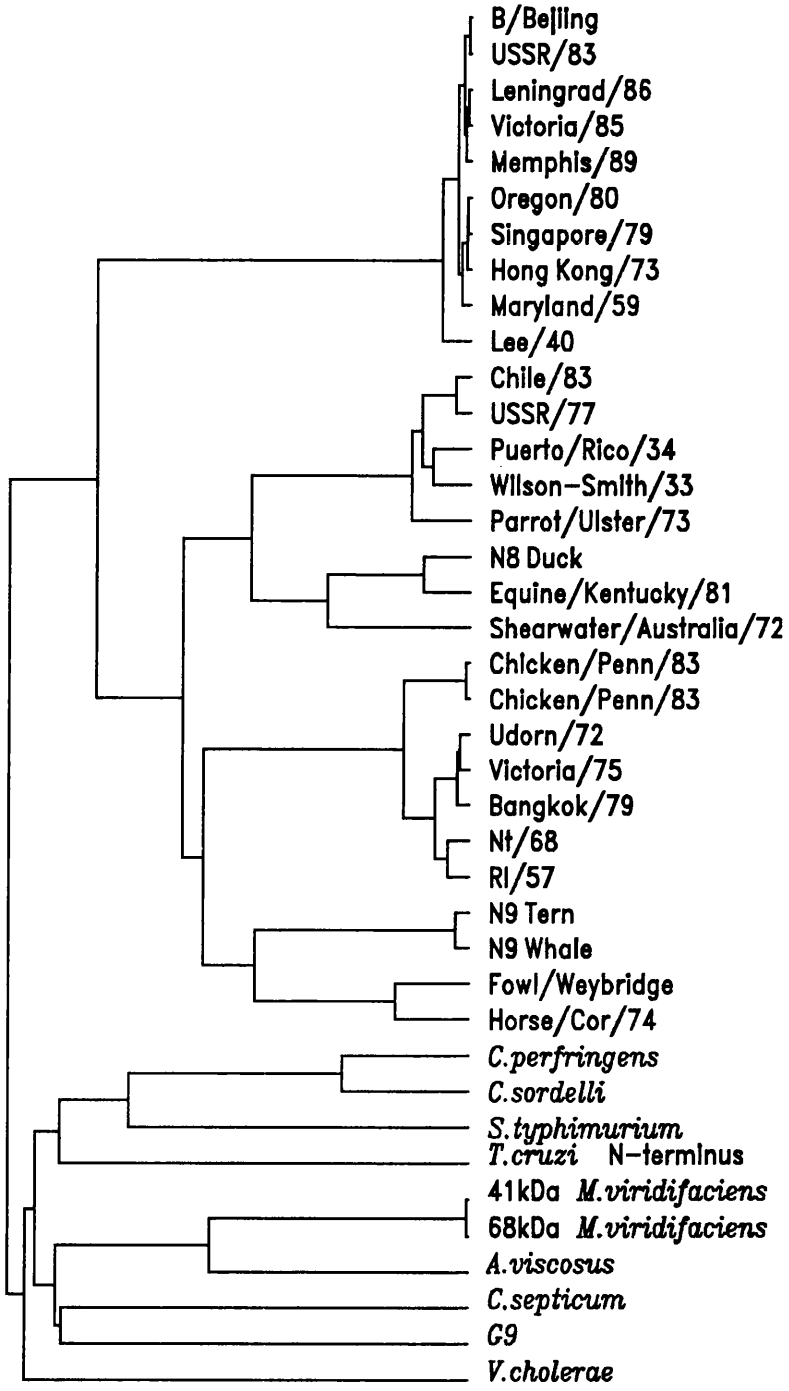
R_2 = N-Acetyl, N-Glycolyl, Amino, Hydroxyl

R_3 = Gal (3/4/6), GalNAc (6), GlcNAc (4/6) or Sialic acid (8/9)
(Absent in 2,6 and 2,7 anhydro compounds)

Sources of neuraminidase

Neuraminidases are ubiquitous in nature and are found in some viruses, pathogenic and non-pathogenic bacteria, some parasites and in all complex metazoans from starfish to humans. The best characterized neuraminidases are those from viral and to a lesser extent bacterial sources. A dendrogram showing the relationship in terms of sequence identity between the viral and non-viral neuraminidases is shown in fig. 1.2. As the grouping indicates, sequence identities are variable between the neuraminidases, being highest between the influenza virus B neuraminidases and lowest between the bacterial neuraminidases (table 1.1).

Fig. 1.2 A dendrogram of the viral and non-viral neuraminidases. Grouping is based on sequence identity only.



(a) Influenza virus neuramindase

Influenza viruses can be divided into two classes, types A and B. Type A can be further subdivided into nine serologically distinct subtypes; N1-N9. Influenza virus neuraminidase preferably cleaves the α2-3 linkage between the terminal sialic acid residue and the penultimate sugar. The role of neuraminidase in the life cycle of the virus is still not completely understood, however two main functions have been assigned to the enzyme. It is well established that there is a link between influenza virus infectivity and neuraminidase activity. Influenza virus neuraminidase aids virus penetration of host epithelial cells by removing sialic acids from mucin and thus lowering the viscosity. The neuraminidase also removes sialic acids from hemagglutinin, the second antigen on the surface of the influenza virus. Hemagglutinin is responsible for the attachment of virus to cells and for the penetration of the virus into the cell during the initial stages of infection. Removal of sialic acids from the hemagglutinin prevents self agglutination of infected host cells (Palese *et al.*, 1974) and allows elution of virus progeny virus particles from infected cells, thus aiding infectivity (Gottschalk, 1966).

The inadequacy of vaccines against influenza virus, due to the rapid antigenic drift and shift that the viral antigens undergo (Murphy & Webster, 1990) has prompted the search for inhibitors against neuraminidase. The first inhibitor to be discovered was a synthetic derivative of sialic acid called Neu5Ac2en (fig. 1.3), or 2-deoxy-2,3-dehydro-N-acetylneuraminic acid (DANA), which is believed to represent a transition state analogue (Meindl & Tuppy, 1969). Many other analogues of DANA have now been synthesized (Meindl *et al.*, 1974; Flashner *et*

al., 1983; Glanzer *et al.*, 1991), but none have K_i values less than micromolar, or show significant activity in vivo. However with the advent of high resolution native and some substrate/inhibitor-bound crystal structures of influenza virus A neuraminidases of subtypes N2 (Varghese *et al.*, 1983) and N9 (Tulip *et al.*, 1991), as well as influenza virus B neuraminidases of strain B/Lee/40 (Janakiraman *et al.*, 1994) and B/Beijing/1/87 (Burmeister *et al.*, 1992), rational drug design has become possible. The rational design of a specific inhibitor of influenza virus neuraminidase will be discussed in Chapter 8.

The structure of influenza virus neuraminidase

The viral neuraminidase is a homo-tetramer of molecular weight 240kDa. It consists of a globular box like-head where the enzymatic activity resides, attached to a hydrophobic stalk which anchors the enzyme in the membrane. The neuraminidase heads can be released by digestion with pronase or trypsin and can be crystallized. Sequence identities range from as low as 25% between subtypes, to 70% or more within a subtype. Identities between viral and non-viral neuraminidases can be as low as 16% (table 1.1). Despite these variable sequence identities, all neuraminidase structures characterized to date possess the canonical neuraminidase β-propeller fold, which consists of six four stranded antiparallel β-pleated sheets wrapped around the blades of a propeller.

Table 1.1 Sequence identities between viral and non-viral neuraminidases.

%	mhc g9	*T.cruzi*	*V.cholerae*	mv[1]	st[2]	N8	N9	N2	B/Beijing	B/Lee
mhc g9[3]	100	21.8	25.7	33.4	28.3	19.1	17.2	17.4	16.7	18.6
T.cruzi	21.8	100	19.9	21.8	31.6	17.4	15.2	15.3	15.1	14.6
V.cholerae	25.7	19.9	100	26.0	26.9	16.9	17.0	19.7	18.0	17.7
mv[1]	33.4	21.8	25.9	100	19.2	16.2	15.4	16.2	18.3	19.8
st[2]	28.3	31.6	26.9	19.2	100	18.2	15.1	16.9	17.5	17.2
N8	19.1	17.4	16.8	16.2	18.2	100	42.3	47.5	33.8	33.6
N9	17.2	15.2	17.0	15.4	15.1	43.0	100	45.6	29.2	29.6
N2	17.4	15.3	19.8	16.2	16.9	47.5	45.6	100	31.6	29.4
B/Beijing	16.7	15.1	18.0	18.3	17.5	33.8	29.2	31.6	100	94.4
B/Lee	18.6	14.6	17.7	19.8	17.2	33.6	29.6	29.4	94.4	100

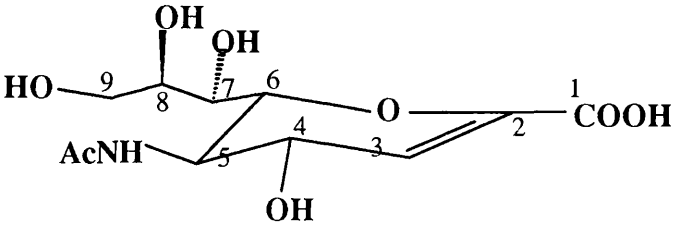
[1] *M.viridifaciens* neuraminidase

[2] *S.typhimurium* neuraminidase

[3] MHC G9 is a neuramindase found in the human major histocompatibility complex (Duncan Campbell and Caroline Milner - personal communication)

Putative calcium ions have been located in the reported X-ray structures of influenza virus neuraminidases. The calcium ion is believed to have a stabilizing effect on the enzyme's active site, although since the presence of calcium is not essential for enzyme activity, it does not have a direct catalytic role (Burmeister *et al.*, 1992).

Fig. 1.3 Neu5Ac2en



(b) The bacterial neuraminidases

Neuraminidases are also widely distributed in bacterial species. Bacterial neuraminidases have varying molecular weights and may be broadly classified into large, or high molecular weight (e.g. *V.cholerae*) and small, or low molecular weight (e.g. *S.typhimurium*). High activities are common and substrate specificity is broader than for the influenza viral neuraminidases, including cleavage of α2-3, α2-6, and less frequently α2-8 linkages. Bacterial neuraminidases are known to have a nutritional role as well as being implicated in pathogenesis.

Structure of the bacterial neuraminidases

The structures of the small bacterial neuraminidase from *S.typhimurium* (Crennell *et al.*, 1993) and the large bacterial neuraminidase from *V.cholerae* (Crennell *et al.*, 1994) have been reported. Despite the low sequence identity exhibited between the bacterial and viral enzymes (~ 15%, table 1.1), both the bacterial neuraminidase structures show a high similarity in the overall fold and in the spatial arrangement of catalytic residues. All the bacterial neuraminidases possess a number of conserved sequence motifs, called Asp-boxes. These have the sequence Ser/Thr-X-Asp-[X]-Gly-X-Thr-Trp/Phe and are found in spatially conserved positions between the third and fourth strands of the β-sheets. The function of Asp-boxes remains unknown, but they may be involved in secretion or protein folding.

The first structure of a bacterial neuraminidase to be solved was that from *S.typhimurium*. *S.typhimurium* neuraminidase is mainly β-sheet, possessing the

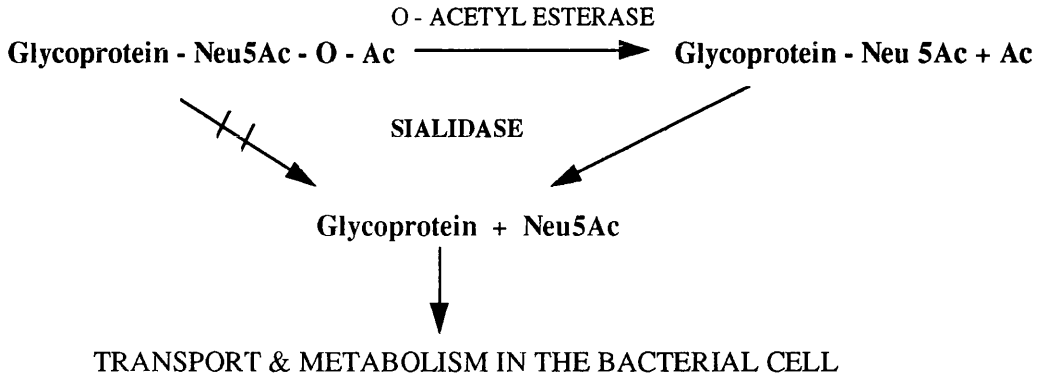
canonical β-propeller fold, and has two small α-helices. There is one disulphide bond between the innermost β-strands of the first and second sheets. Although the fold of the enzyme is identical to that of the influenza viral neuraminidases, there are major differences in the lengths of the loops and β-strands. The neuraminidase from *S.typhimurium* possesses four conserved Asp-box motifs in between the third and fourth strands of the first four sheets.

The second structure of a bacterial neuraminidase to be reported was that from *V.cholerae*. The neuraminidase from *V.cholerae* belongs to the class of large bacterial neuraminidases and has a molecular weight of 83kDa. Again the enzyme is mostly β-sheet, with 5 small α-helices. In addition to the typical canonical β-propeller neuraminidase domain, the enzyme also has two lectin-like domains (wings) which flank the former domain. Each lectin-like domain consists of a sandwich formed from two sheets of seven and six antiparallel β-strands, having a topology most like that of the legume lectins. The purpose of these wings in the pathogenicity or life-cycle of *V.cholerae* is unknown, and the wings do not retain any of the carbohydrate binding residues found in the lectins. However the carbohydrate binding sites of the lectins are located on the concave surface of the fold and the equivalent surfaces in *V.cholerae* neuraminidase are solvent exposed, thus suggesting that the wings do recognize a ligand (Crennell *et al.*, 1994). *V.cholerae* neuraminidase has four Asp-boxes, together with a further Asp-box which deviates from the consensus sequence. In contrast to the small bacterial neuraminidases the enzyme requires Ca^{2+} for activity.

Non-pathogenic neuraminidases

The terminal non-reducing position of sialic acids in sialo-glycoconjugates requires that the sialic acid must be removed before metabolism of the remaining oligosaccharide chain can occur. Many non-pathogenic neuraminidases function in symbiotic relationships at mucosal surfaces, such as the oral cavity and the human large intestine. For some neuraminidases, no role has yet been defined, e.g. the soil bacteria *Arthrobacter* sp. and *M.viridifaciens*. However they may simply fulfil part of the nutritional requirement of the bacterium, metabolizing the sialic acid as a carbon and energy source. In the human large intestine, these bacteria belong to the enteric bacteria and comprise the most abundant anaerobes, e.g. *Bacteriodes* sp., facultative anaerobes such as *E.coli* and *Enterococcus faecalis* strains, and a group of mucin oligosaccharide-degrading bacteria including *Ruminococcus* and *Bifidobacterium* sp. which are capable of completely degrading the glycoproteins present in the mucus at the surface of the normal mucosa (Hoskins *et al.*, 1985). A role for these enzymes in providing an energy source for the bacteria through mucus degradation has been demonstrated (Hoskins *et al.*, 1985). However a balance is required between synthesis of the mucin, which provides a defence mechanism and the degradation to provide a source of energy. O-acetylation of sialic acids is known to retard the rate of action of bacterial neuraminidases (Corfield *et al.*,1986, 1992) and colonic mucins contain a typically high sialic acid O-acetylation pattern. In addition, the presence of an O-acetyl sialic acid esterase in many human enteric bacterial strains provides additional regulation. A scheme depicting this method of regulation of sialic acid degradation in human colonic bacteria is illustrated in fig 1.4.

Fig.1.4 Regulation of sialic metabolism in human colonic bacteria. More than two O-acetyl esters on a sialic acid results in retardation of the rate of catalysis by the neuraminidase. A specific O-acetyl esterase removes the acetyl esters, thus facilitating neuraminidase action.



A further example of the nutritional role of neuraminidases has been shown for *S.typhimurium* LT2 (Vimr & Hoyer, 1992). The nan H gene coding for the neuraminidase was only found in the subgenus 1 isolate for LT2. Thus, the presence of the neuraminidase is not necessary for pathogenesis in this genus (Hoyer *et al.*, 1992). A *S.typhimurium* mutant, deficient in neuraminidase was used to compare the growth of the wild type and mutant strains using sialyl α(2,3)-lactose as the sole carbon source. *Salmonella* strains cannot use lactose as a carbon source and sialic acid must be released first before it can be taken up and metabolized by the bacteria. As a result, the wild type bacterium was able to grow on the sialyl-lactose medium, but the neuraminidase negative mutant was not (Vimr & Hoyer, 1992).

Neuraminidases associated with pathogenicity

Neuraminidase is one of the many hydrolytic enzymes often associated with bacterial invasion of the host. Some of the diseases associated with neuraminidase activity, such as meningitis and glomerulonephritis are localized to defined sites where they cause damage. Others such as gas gangrene and peritonitis, exhibit a much wider pattern of colonization and bacteria are found regularly in the circulation. A summary of the more important infectious diseases associated with neuraminidase activity are listed in table 1.2.

Table 1.2 Disease related neuraminidases and their major site(s) of infection.

Disease	**Bacteria**	**Site**	**Found in blood?**
Gas gangrene	Clostridia	wounds	yes
Septicaemia	Streptococcus Pneumococcus Bacteriodes Corynebacteria	various	yes
Pneumonia	Streptococcus	respiratory tract	yes
Peritonitis	Clostridia Bacteriodes Enterococcus Escherichia other intestinal bacteria	peritoneum	yes
Meningitis	Streptococcus Group B type III	brain	yes
Haemolytic-uraemic syndrome	Pneumococcus	kidney	yes
Glomerulonephritis	Streptococcus type A	kidney	yes
Periodontal disease	Streptococcus Haemophilus Actinomyces Other oral bacteria	oral cavity	no
Otitis media	Streptococcus	middle ear	no
Cholera	Vibrio cholerae	gastrointestinal tract	no

Because of the involvement of sialic acids in cellular functions, enzymic removal by neuraminidase leads to marked differences in the biological behaviour of cells and molecules; sometimes resulting in pathogenesis. Five main functions for sialic acids can be summarized:

1. As a result of their strong negative charge, sialic acids are involved in the binding and transport of positively charged compounds and in aggregation via Ca^{2+} bridges, or disaggregation by repulsive effects in cells.

2. Sialic acids influence the conformation of glycoproteins, which is important for the correct arrangement of glycoprotein molecules in cell membranes, for the resistance of glycoproteins to proteases and for the activity of glycoprotein enzymes.

3. N- and O- acetylated sialic acids are antigenic determinants in various mammalian and microbial systems (Schauer, 1983). In addition, sialic acids contribute to the specificity of blood group substances (Roelcke, 1984).

4. Sialic acids are essential components of receptors, e.g. for peptide hormones, toxins and viruses (Paulson *et al.*, 1984). Infection of cells by these microorganisms is exclusively dependent on the presence of sialic acids on cell membranes.

5. Sialic acids prevent recognition of receptors by the corresponding ligands, or vice versa, and recognition of antigenic sites by components of the immune defence system. The role of sialic acids as biological masks is possibly the most important role of sialic acids (Schauer *et al.*, 1984) and will be discussed in more detail.

Removal of biological masks by neuraminidases associated with pathogenicity

A glycoprotein layer rich in sialic acids acts as an immunobarrier between mother and embryo. Suppression of antigenicity by sialic acids appears to be a general feature of molecules and cells and partial loss of these sugars has been proposed as a cause of autoimmune disease (Schauer, 1982; Reutter *et al.*, 1982; Kottgen *et al.*, 1979). Tumour cells have often been observed with highly sialyated surfaces and are thus protected from immune defence. This may contribute to the spread of the tumour. There are numerous other examples of sialic acids acting as biological masks, the removal of which can lead to various disease states. These include decreasing the reactivities of IgG with the Fc receptor of human T lymphocytes; cytotoxic complement-dependent factors against autologous, invasive human bladder tumour cells; killer cells; some receptors for hormones or drugs; and the binding of rabbit blood platelets to endothelial cells (Schauer *et al.*, 1984). Two examples of disease related to neuraminidase action are discussed in detail:

1. Cholera. *V.cholerae* produces a potent neuraminidase which may play a role in the pathogenicity of the bacterium. Gangliosides are carbohydrate-rich sphingolipids that contain at least one sialic acid. The oligosaccharide chain is linked to ceramide by a glucose residue. The action of *V.cholerae* on cell surface gangliosides yields G_{M1} gangliosides as a product because the galactose-linked sialic acid cannot be released by the enzyme (Venerando *et al.*, 1982). G_{M1} gangliosides consist of a galactose β-1,3 linked to N-acetylgalactosamine, which is β-1,4 linked to a further galactose moiety, which in turn is α-2,3 linked to N-acetyl neuraminic acid and β-1,4 linked to glucose. The oligosaccharide chain ends with

ceramide linked via the glucose. G_{M1} then acts as a receptor for the cholera enterotoxin at the mucosal cell surface, facilitating its penetration (Muller, 1976; Galen *et al.*, 1992).

2. Septicaemia. Strains of *Streptococcus*, pneumococcus, corynebacteria and *Bacteriodes* are amongst the many neuraminidase-producing bacteria which are primary agents in septicaemia. Analysis of the serum glycoproteins from patients with septicaemia due to *Streptococcus sanguis* strains showed that sialic acid had been lost. However the association of virulence with neuraminidase activity remains unclear. No requirement of neuraminidase as a virulence factor has been demonstrated, although a high-virulence clone of group B type III *Streptococcus* has been identified which produces elevated levels of extracellular neuraminidase (Musser *et al.*, 1989).

In summary, individual examples of a nutritional role, as in *S.typhimurium* LT2 (Vimr & Hoyer, 1992), or the absence of a nutritional role, as in *P.aeruginosa* (Cacalano *et al.*, 1992), may imply that combined roles for neuraminidase are likely to apply elsewhere. In this way, the bacteria can create an energy source, can overcome host defensive mechanisms and can create binding sites for colonization. A nutritional role may thus be considered as a real function of neuraminidase in both pathogenic and non-pathogenic bacteria.

(c) Mammalian neuraminidases

Neuraminidases from mammalian sources have been implicated in several important metabolic processes, such as the catabolism of gangliosides and

glycoproteins (Tulsiani & Carubelli, 1970; Usuki *et al.*, 1988a), the developmental modelling of myelin (Saito & Yu, 1992), the regulation of cell proliferation (Usuki *et al.*, 1988b) and the clearance of glycoproteins from the serum (Ashwell & Morell, 1974). An important potential role for endogenous neuraminidases is in regulation of cell surface sialyation. It has been well documented that how a cell is seen by its external environment is determined in a large part by cell surface sialic acid (Whiteheart *et al.*, 1990). There is indirect evidence to support this role for cellular neuraminidases. A decrease in sialic acid with no consistent alterations in sialyltransferase has been found in Rous sarcoma virus-transformed chick embryo fibroblasts (Spataro *et al.*, 1975). Among various other enzymes involved in the turnover of sialic acids, neuraminidase was the only enzyme whose activity was altered in the virally transformed cells and the changes were correlated with the decrease in cellular sialic acid content. An increased level of neuraminidase activity has also been frequently observed in association with a decreased extent of cell surface sialyation in response to drug treatment (Sodhi & Prasab, 1985), during cell differentiation (Nojiri *et al.*, 1982), or activation (Landolfi *et al.*, 1985; Taira & Nariuchi, 1988). However whether these changes in neuraminidase activity are instrumental in the reported fluctuations in cell surface sialyation remains to be ascertained. Most cellular forms of the enzyme have not been well characterized at the molecular level and at present there is no crystal structure for a mammalian neuraminidase. This is due, in part, to the relatively low amounts found in all tissues and cell types. In addition, many forms of mammalian neuraminidase are extremely labile or tightly membrane associated. The enzyme has been localized in the lysosomal matrix (Tulsiani & Carubelli, 1970; Miyagi *et al.*, 1990), the lysosomal

membrane (Miyagi & Tsuiki, 1984; Zeigler *et al.*, 1989; Sagawa *et al.*, 1990), the plasma membrane (Sagawa *et al.*, 1990), the myelin membrane (Saito *et al.*, 1992) and the cytosol (Miyagi & Tsuiki, 1985; Tulsiani & Carubelli, 1970). To date, at least five distinct mammalian enzyme forms have been identified, each with different kinetic properties and substrate specificities. The enzyme from Chinese hamster ovary cells has been cloned, sequenced and expressed (Ferrari *et al.*, 1994) and the amino acid sequence deduced. The amino acid sequence has no significant identity with any other mammalian protein, or indeed other microbial neuramindases. Since the bacterial neuraminidases have low sequence identity (table 1.1), in the order of only ~30%, this is not surprising. The mammalian sequence does however contain two Asp-box motifs. As detailed in Chapter 8, Asp-boxes have the consensus sequence Ser/Thr-X-Asp-[X]-Gly-X-Thr-Trp/Phe, are of unknown function and appear to be conserved across the bacterial neuraminidases.

(d) The Trans-sialidase (neuraminidase) family of *Trypanosoma Cruzi*

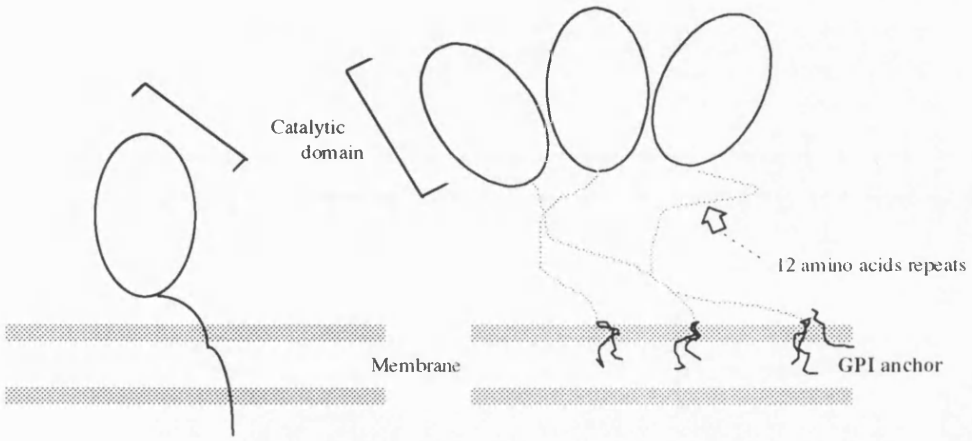
Protozoan parasites of the genus Trypanosoma express a unique sialic acid metabolising enzyme, called a trans-sialidase. Trans-sialidase catalyses the transfer of sialic acids from host glycoconjugates to receptor molecules of the parasite plasma membrane. The preferred sialic acid donors are sialyl α-2,3 linked to β-galactosyl residues on glycoproteins, glycolipids and oligosaccharides (Vandekerckhove *et al.*, 1992). The enzyme is expressed during infectious stages of the life cycle of the human pathogen *Trypanosoma cruzi* and in the procyclic forms of *Trypanosoma brucei*. The enzyme from both trypomastigotes is the

subject of much interest, due to the pathogenic nature of the organisms. *T.cruzi* is the causative agent of Chagas' disease in Central and South America. *T.brucei* is the causative agent of sleeping sickness in man and nagana in game and domestic animals in Africa. However the trans-sialidase is only found in procyclic trypomastigotes of *T.brucei* and not in the mammalian stage bloodstream form.

Trans-sialidase structure

The native and purified enzyme forms multimeric aggregates of molecular mass >400kDa that migrate as multiple bands ranging from 100 to 220kDa in SDS-PAGE upon denaturation (Pereira *et al.*, 1991; Schenkman *et al.*, 1992). The bands differ in terms of their number and relative molecular weight depending on the strain of the parasite (Pereira *et al.*, 1991). The enzyme is linked to the membrane in the parasite by a glycosyl phosphatidylinositol (GPI) anchor. The enzyme is released from trypanosomes by bacterial phospholipase C (Jazin *et al.*, 1991; Rosenberg *et al.*, 1991). Native trans-sialidase is formed by two separate domains, a conserved segment in the N-terminus containing the catalytic site and a heterogeneous region in the C-terminus necessary for enzyme oligomerization or aggregation (fig. 1.5).

Fig. 1.5 Model of *T.cruzi* epimastigote (a) and trypmastigote (b). The epimastigote form of the trans-sialidase is monomeric, whereas the trypmastigote is multimeric. The two foms possess a similar catalytic domain, but the trypmastigote enzyme has an additional domain at the carboxy terminus composed of the 12 amino acid repeats responsible for oligomerization.



(a) Epimastigote trans-sialidase (b) Trypomastigote trans-sialidase

All members of this family characterized so far have retained in the same positions three conserved Asp-boxes and a Fn3 domain containing a degenerate Asp-box and a highly conserved sequence of unknown function, VTV X NV X LYNR, located downstream of the catalytic domain.

At present there is no crystal structure for a trans-sialidase. However the *T.cruzi* trans-sialidase has a sequence identity of 31% with the neuraminidase from *S.typhimurium*, for which the crystal structure is known (Crennell *et al.*, 1993). In addition, the catalytic residues identified in the *S.typhimurium* enzyme are conserved in the trans-sialidase, as are the positions of the first three Asp-boxes.

It has been suggested that the trans-sialidase aids invasion of the target cell by the enzyme possessing a lectin-like activity towards the cellular and extracellular sialic acids. The enzyme will then function as a poor neuraminidase and the parasite may bind to rather than remove the host cell sialic acid. In support of this idea, the purified trans-sialidase inhibits cell invasion when added to the incubation medium (Ming *et al.*, 1993).

To fully resolve many of the unanswered questions regarding the Trypanosoma trans-sialidases such as catalytic mechanism, a high resolution crystal structure will undoubtedly be necessary.

(e) Hemagglutinin neuraminidase

The paramyxoviruses are a group of negative stranded, enveloped, RNA-containing viruses that include Newcastle disease virus, mumps virus and the parainfluenza viruses (including Sendai virus, parainfluenza virus 3 and simian virus 5) (Kingsbury *et al.*, 1978). There are two types of spikes protruding from the virion membrane of paramyxoviruses. These are the fusion protein (F) and hemagglutinin neuraminidase (HN) glycoprotein spikes. Unlike the influenza virus, where the neuraminidase and hemagglutinin activities are carried by separate glycoproteins, the paramyxoviridae carry a bifunctional molecule; hemagglutinin neuraminidase.

The Dual functions of hemagglutinin neuraminidase

Hemagglutinin neuraminidase carries out the opposite functions of attachment to sialic acid containing cellular receptors and release from the same

moiety, catalysed by the neuraminidase activity associated with the protein (Choppin & Scheid, 1980). Following exposure of a hydrophobic sequence by proteolytic cleavage, the F protein mediates both virus-cell and cell-cell fusion (Hsu *et al.*, 1981; Scheid & Choppin, 1974). For many of the paramyxoviruses, including Newcastle disease virus, participation of the hemagglutinin neuraminidase spike is also necessary for membrane fusion. The hemagglutinin neuraminidase mediates attachment to cellular receptors and it has been shown that mutations in the hemagglutinin neuraminidase that influence the strength of its interaction with cellular receptors also influence the extent of syncytium formation (Iorio & Glickman, 1992).

Structural studies of hemagglutinin neuraminidase

At present there is no three-dimensional structure available for a hemagglutinin neuraminidase. However, considerable information is available about structure-function relationships of the enzyme in this genus. Ultrastructural and biochemical evidence exists that Sendai virus hemagglutinin neuraminidase molecules exist as disulphide-bonded dimers and tetramers on the virion surface (Markwell & Fox, 1980; Thompson *et al.*, 1988). These homomultimers will presumably be able to carry out their functions with greater efficiency than could monomers. Extensive comparisons of the predicted amino acid sequences of a number of hemagglutinin neuraminidases have now been made following the elucidation of the primary sequences of the enzyme from representatives including Newcastle disease virus, mumps virus, Sendai virus and parainfluenza virus (Morrison & Portner, 1991). Colman *et al.* have proposed a model for the

hemagglutinin neuraminidase three dimensional structure (Colman *et al.*, 1993). This model is based on the identification of invariant amino acids among hemagglutinin neuraminidases which have counterparts in the enzyme active site of influenza virus neuraminidase. The model is broadly similar to that of the influenza virus neuraminidase and has the canonical β-propeller fold.

The Enzyme in question

Of the genus *Micromonospora*, *M.viridifaciens* shows the highest neuraminidase enzymatic activity (Aisaka & Uwajima, 1987). The neuraminidase from *M.viridifaciens* is a typical inducible enzyme. When colominic acid is used as an inducer, a soluble 41kDa neuraminidase is secreted, however when milk casein is used a higher molecular weight 68kDa soluble neuraminidase is secreted (Sakurada *et al.*, 1992). Neuraminidase from *M.viridifaciens* is monomeric, as with the other bacterial neuraminidases that have been characterized and possesses 5 typical Asp-boxes. The neuraminidase MHC G9, a human enzyme from the major histocompatibility complex has highest sequence identity with *M.viridifaciens* neuraminidase (table 1.1), although the neuraminidase from *C.perfringens* has many similarities with *M.viridifaciens* neuraminidase. Both have similar methods of expression regulation, some catalytic properties are similar and the deduced processing site of *M.viridifaciens* neuraminidase agrees with the C-terminal end of *C.perfringens* neuraminidase when the two sequences are aligned by their homology (Aisaka *et al.*, 1991; Nees *et al.*, 1975). The structures of neuraminidases from non-pathogenic bacteria, including *M.viridifaciens*, have not been characterized until now. The enzyme has a broad range of specificity with a

high activity; in the range of 200 units/mg protein (Aisaka *et al.*, 1991); which is comparable to that for *C.perfringens* (150-500 units/mg) (Cassidy *et al.*, 1965; Bouwstra *et al.*, 1987; Drzeniek, 1972) and *V.cholerae* (133 units/mg) (Drzeniek, 1972). The enzyme has its maximum activity at pH 5.0 when colominic acid or NANA-lactose was used as substrate (pH optimum for bacterial neuraminidases lies in the range 4.5-6.5 (Drzeniek *et al.*, 1972)). The neuraminidase activity has previously been shown to reside in the N-terminal 41kDa portion of the enzyme (Sakurada *et al.*, 1992), however the function of the remaining 27kDa is unknown. The precise role of the enzyme in cell metabolism remains to be established. Table 1.3 summarizes some of the differences between the neuraminidase from *M.viridifaciens* and other microbial species.

Table 1.3 Comparison of the properties of some microbial neuraminidases. Molecular weights were determined by three methods: [a] by gel filtration [b] by SDS-PAGE and [c] from the DNA sequence.

Organism	Molecular Weight	K_m for NANA-lactose (mM)	Ca^{2+} requirement?	Inhibitors
Actinomycetes				
M.viridifaciens	41,000[b] 68,000[b]	2.1	no	PCMB,NBS, I_2
Streptomyces albus	40,000-45,000[a]	0.45	no	
Streptomyces griseus	32,000[a]	0.15-0.40	no	Hg^{2+}, Fe^{2+}
Other Bacteria				
Arthrobacter I	51,000[a]	0.6	no	NBS, I_2
ureafaciens II	39,000[a]	0.6	no	NBS, I_2
Bacteriodes fragilis	92,000[a]	0.51	no	
Clostridium pefringens	42,800[c] 56,000-58,000[a] 65,000-67,000[a,b]	0.8-2.4 1.2-1.6	no	PCMB, Hg^{2+}, Fe^{2+}
Corynebacterium diphtheriae	65,000[a]	0.9	yes	EDTA
Streptococcus group A	90,000[a]	no activity	yes	
Streptococcus group B	106,000[b] 125,000[a]	no activity no activity	yes yes	
V.cholerae	68,000[a] 90,000[a]	1.2 1.0	yes yes	EDTA EDTA

The Organism - *Micromonospora viridifaciens*

The neuraminidase studied in this project is produced by the non-pathogenic actinomycete *Micromonospora viridifaciens*. Actinomycetes are a group of unicellular gram-positive bacteria that are characterized by the formation of branching filaments. Micromonosporas occur both in soil and in fresh water, but

the majority are found in the latter. Both of these members of the genus are strict anaerobes and many of them can utilize polysaccharides such as cellulose, chitin and xylan as carbon energy sources. Anaerobic, cellulose-fermenting species of *Micromonospora* have been isolated from the gut contents of termites and are the only members of this group capable of fermentation. Micromonosporas are members of the actinomycetes known as *euactinomycetes*. These develop only in the mycelial state and they reproduce through the formation of unicellular spores.

The *Micromonospora* strains are of significant industrial interest, due to their ability to produce antibiotics. Antibiotics produced by this genus include gentamicins (Maehr & Schaffner, 1967), sagamicins (Okachi *et al.*, 1974), sisomicins (Weinstein *et al.*, 1970) and fortimicin A (astromicin) (Nara *et al.*, 1977). As a consequence an efficient gene cloning strategy has been developed for *Micromonospora* strains (Hasegawa *et al.*, 1991). However despite the interest in this genus, relatively little is known about the physiological and biochemical properties of *Micromonospora* strains, in particular *M.viridifaciens*.

Aims of the project

The aim of this project was to solve the three dimensional structures of both the 41kDa and 68kDa forms of the neuraminidase from the bacterium *Micromonospora viridifaciens*. Ideally the structure of the enzyme-inhibitor complex was also to be solved in order that detailed structural comparisons could be made between the active site of the enzyme and other neuraminidases.

Chapter 2

PURIFICATION AND CRYSTALLIZATION

Introduction

In order to obtain protein crystals which are of as high a quality as possible, much effort must be put into purifying the protein in question to a high degree of homogeneity. Protein impurity is often a cause for the failure of proteins to crystallize successfully (Abergel *et al.*, 1991). The purification strategy for both forms of the enzyme was based on a protocol developed by Taylor *et al.*, (1992), utilizing fplc.

Obtaining protein crystals suitable for X-ray diffraction studies is still one of the major rate determining steps in the solution of a crystal structure and the technique of crystallization is still regarded by many as an art rather than a science. However, if a high resolution crystal structure is to be obtained, one must persist in their efforts to achieve good quality protein crystals, which diffract to a high resolution. Firstly the crystals should be large, since the intensity of the diffraction pattern is directly proportional to the volume of the crystal within the limits of absorption effects and secondly they should be fairly robust, so as to allow extensive data collection without rapid degradation of the crystal due to X-ray damage.

However at present, we have a relatively poor understanding of the phenomena surrounding macromolecular crystallization and therefore we must resort to more empirical methods. This involves screening as many different conditions as possible to search for the optimum crystallization condition for a given protein. However the number of possible solutions to be tested is so great that an exhaustive search would be impractical (McPherson, 1982). Hence a sparse matrix of screening conditions, which samples a large range of buffers, pH, additives and precipitants is often used initially (Jancarik & Kim, 1991) to find approximate crystallization conditions.

Crystal growth

Crystals grow from a supersaturated solution, since the system must be in a non-equilibrium to provide the thermodynamic driving force for crystallization. In order to achieve the limited degree of supersaturation required for crystallization, several objectives must be met. The first is to deprive the protein molecules of enough water molecules to disrupt their hydration shell; whilst decreasing the effective electrostatic shielding between the macromolecules. These factors tend to induce phase separation and force the system towards a new free energy minimum. This maximizes attractive interactions and if the point of inadequate solvation is reached slowly enough the macromolecules are allowed sufficient opportunity to order themselves in a crystalline lattice. Modification of the solvent system, such as by the addition of precipitants, can encourage the process of supersaturation by guiding the system very slowly towards a state of minimum solubility. However

due to the complex and rather unpredictable behaviour of macromolecules, as a result of their polyvalent surface character, varied shapes and dynamic properties, the correct degree of supersaturation for a given protein is not always easy to achieve. If a protein does prove difficult to crystallize, then a likely cause may be protein impurity, or microheterogeneity in the protein sample. Since crystals are essentially composed of perfect symmetry and periodic translational relationships between the molecules in the lattice, the introduction of impurities will result in inhibition of crystal growth since the non-uniform protein units will not be able to enter the crystal properly and will not bear proper correspondence to their neighbours. If early inhibition of crystal growth does present a problem, then a possible method of overcoming this is by the technique of seeding. Macroseeding involves transferring the crystal from the initial drop through a series of drops not containing protein, each containing increasing concentrations of precipitant. With successive transfers, any impurities on the surface of the crystal that are preventing continued crystal growth are hopefully removed and the crystal can then be transferred to a fresh drop containing protein and a decreased concentration of precipitant relative to the previous drop, determined by experience. Microseeding involves removing microseeds from a trial which may then provide nucleation sites for further trials. The microseeds are aspirated and then simply transferred to a fresh drop.

Crystallization techniques

There are various methods employed in the crystallization of biological macromolecules, including the growth of crystals in gels, the use of microdialysis, sitting drop and hanging drop vapour diffusion methods. The same basic principles outlined above are employed in all cases, and varying degrees of success have been recorded. Gel matrices have been found to suppress nucleation, reduce the rate of growth and generally lead to higher quality crystals with enhanced stability (Cudney *et al.*, 1994). Microdialysis has the advantage that by liquid-liquid diffusion through a semi-permeable membrane, a protein solution can be exposed to potential crystal producing conditions without actually altering directly the mother liquor. In addition, because the rate of change of substituents in the mother liquor is proportional to the gradient of concentrations across the membrane, the nearer the system approaches equilibrium, the more slowly it changes (McPherson, 1989). However, the vapour diffusion methods are still the most widely employed crystallization techniques today. Hanging drop vapour diffusion was the method implemented here, and as the name implies, uses vapour phase equilibration. A droplet of mother liquor, mixed with a droplet of the protein sample, is placed on a siliconized cover slip, inverted and placed over a tissue culture well, containing the precipitating solution. The cover slip is sealed with silicone vacuum grease. The transport of water through the vapour phase between the microdrop and the reservoir results in the protein concentration in the microdrop slowly coming to super saturation. Seeding techniques in the hanging drop are outlined as in the section on Crystal Growth.

Purification of 41kDa *M.viridifaciens* neuraminidase

Crystals of the 41kDa *M.viridifaciens* neuraminidase had previously been grown by Elizabeth Dineley from protein purified from a culture filtrate of *M.viridifaciens*. The gene for the enzyme was later cloned (Sakurada *et al.*, 1992) and the crude recombinant protein in a freeze-dried form (a generous gift from Mamoru Hasegawa, Kyowa Kogyo Co. Ltd., Tokyo, Japan) was used as the starting material in order to repeat the purification and obtain further crystals.

The recombinant protein was first dissolved in 10mM Tris-HCL buffer, pH 8.0. This sample was run down an fplc Mono-Q anion exchange column, also equilibrated with 10mM Tris-HCL pH 8.0 and eluted with a NaCl gradient from 0 to 0.2M. 1ml fractions judged to contain only/mostly the 41kDa protein by SDS PAGE (Laemmli, 1970) were pooled and then dialysed extensively against 10mM Tris-HCL pH 8.0.

The dialysed sample was then concentrated by spinning through a microsep concentrator (30K cut-off) until the desired volume was reached to give a protein concentration of approximately 10mgml^{-1}. The concentrate was determined to be nearly 100% pure by SDS-PAGE (Fig. 2.1). A typical purification table is shown in Table 2.1.

Purification of 68kDa *M.viridifaciens* neuraminidase

The crude protein used as the starting point for the purification was an ammonium sulphate precipitation extract (kindly donated by Dr T. Uwajima, formerly of Kyowa Kogyo Co. Ltd., Tokyo, Japan), which was dissolved in 10mM

Tris-HCL at pH 8.0. The sample was then centrifuged at 6,000 rpm for about 10 minutes and the supernatant dialysed overnight against 10mM Tris-HCL pH 8.0, to ensure that the sample was salt-free. After concentration to a suitable volume, the dialysed sample was run down an fplc Mono-Q anion exchange column, again equilibrated with 10mM Tris-HCL pH 8.0 and fractions were eluted using a 0 to 0.4M NaCl gradient. Fractions were selected for homogeneity of the 68kDa component by SDS-PAGE, pooled and run down a fplc Superdex 200 gel filtration column equilibrated with the same buffer. Appropriate fractions were pooled and concentrated and SDS-PAGE showed the degree of purity of the sample (Fig. 2.1). The purification table is shown in Table 2.2.

Table 2.1 Purification of 41kDa *M.viridifaciens* neuraminidase, using anion exchange chromatography.

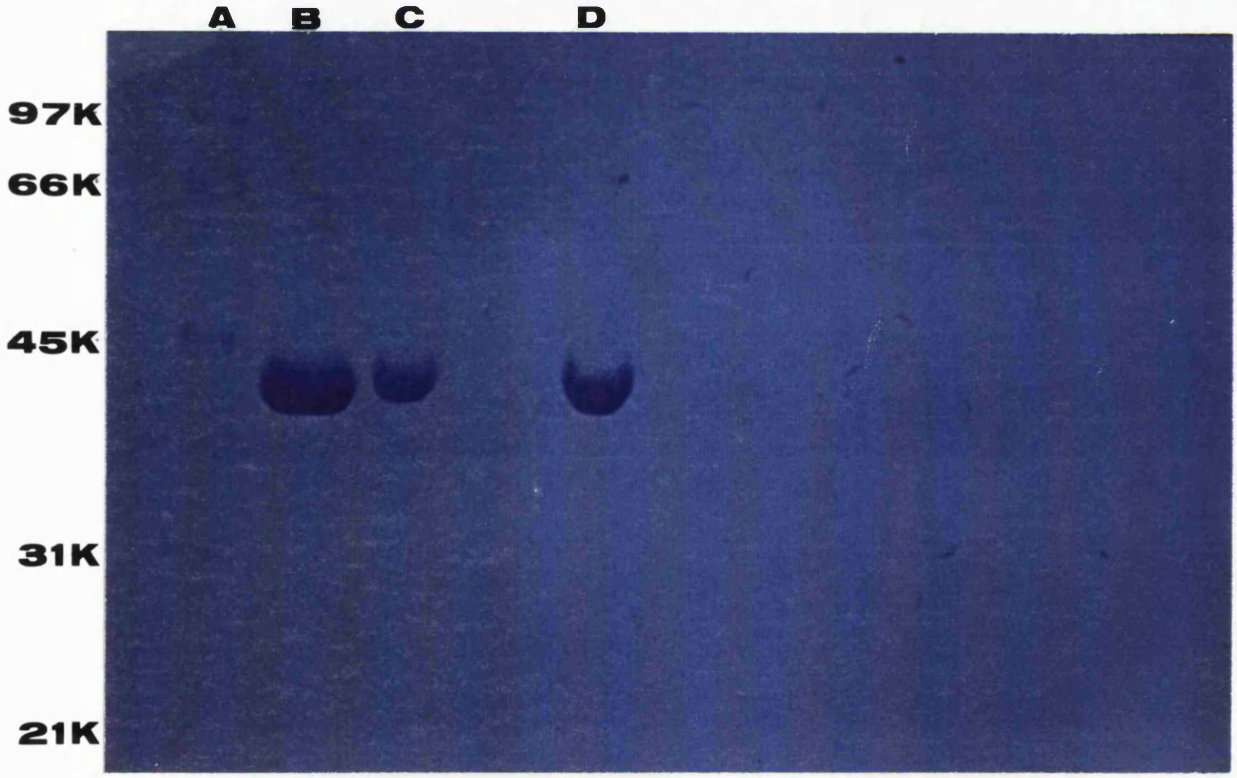
step	total protein (mg)	total activity (nmol/min)	specific activity (nmol/min/mg)	purification (fold)	yield (%)
crude	5.0	131000	26200	1	100
mono-q	1.8	106560	59200	2.26	81

Table 2.2 Purification of 68kDa *M.viridifaciens* neuraminidase, using anion exchange and gel filtration chromatography.

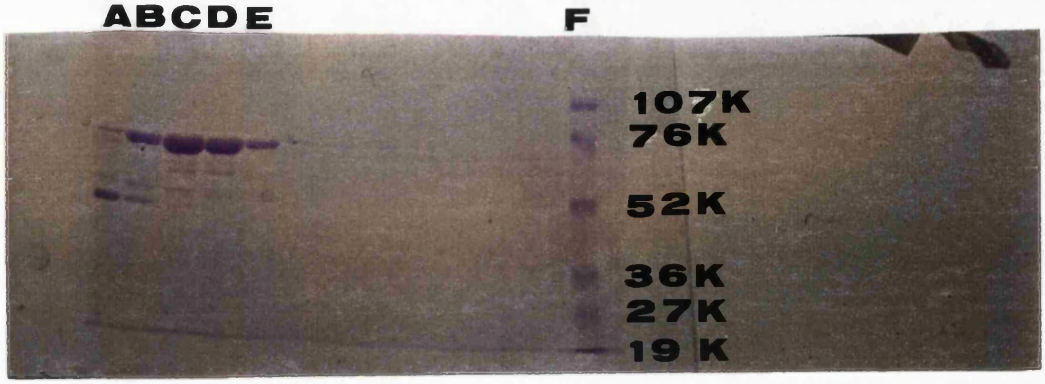
step	total protein (mg)	total activity (nmol/min)	specific activity (nmol/min/mg)	purification (fold)	yield (%)
dialysed crude sample	12.56	204979	16320	1	100
mono-q	2.54	143383	56450	3.46	70
superdex	1.49	87016	58400	3.58	42

Fig. 2.1 10 % SDS-PAGE gels displaying the purity of **a)** 41kDa protein, **b)** 68kDa protein.

a) SDS-PAGE of the 41kDa protein after fplc mono-q anion exchange. Lane A shows the molecular weight markers. The same purified 41kDa protein sample was applied to lanes B, C and D, but the concentration of the sample applied to lane B (approximately 10mgml $^{-1}$) was twice that of C and D.



b) SDS-PAGE of protein fractions that eluted from the superdex gel filtration column (lanes A to E). The first major band in lanes A to E corresponding to the purified 68kDa protein is clearly visible. Lane F shows the molecular weight markers.



Crystallization trials

All trials were set up using the hanging drop vapour diffusion method. Protein concentration and volume in the drop, type and pH of buffer, molecular weight and concentration of the precipitant PEG and volume of the reservoir solution were all varied to find the optimum conditions for crystallization.

Crystallization of the 41kDa form of *M.viridifaciens* neuraminidase

The optimum conditions found for crystallization of the 41kDa form were based on those previously reported by Taylor *et al.* (1992). A variety of conditions gave good quality crystals, these being based on 8% PEG 3350 as precipitant at pH 5.0 to 5.5. Drop sizes were 6µl, consisting of equal volumes of reservoir solution and protein of concentration 12mg/ml. Large crystals were observed, with

dimensions up to 1.4mm x 0.5mm x 0.5mm (fig. 2.2), after about a week of incubation at room temperature.

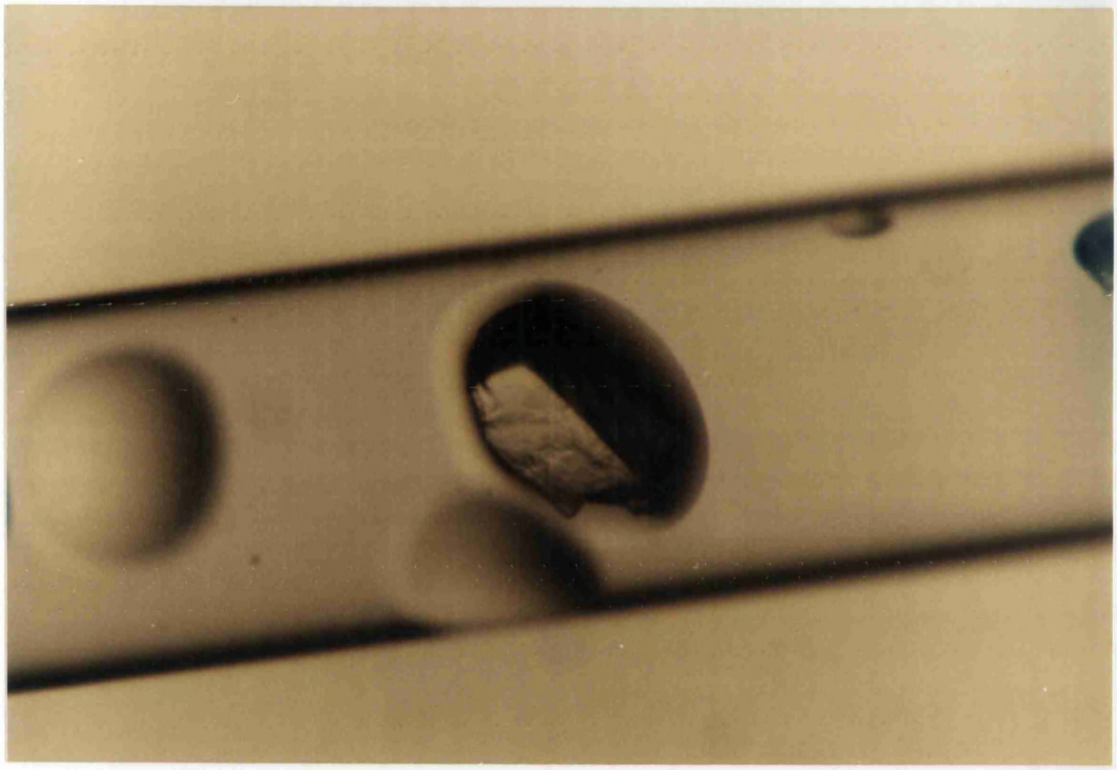
Fig. 2.2 Typical 41kDa crystals.

Crystallization of the 68kDa form of *M.viridifaciens* neuraminidase

In comparison with the 41kDa crystals, good quailty 68kDa crystals were more difficult to obtain. Initially, small crystals were obtained of dimension 0.23 x 0.23 x 0.23mm, using 2% PEG as precipitant, pH=5.0, with a protein concentration of 8mg/ml and total dropsize=8µl. These crystals diffracted weakly to 6Å. After an exhaustive screening process, including use of the magic 100 incomplete factorial matrix (Jancarik & Kim, 1991) which was however unsuccessful, crystals were

obtained of a different crystal type as judged by subsequent analysis, by microseeding followed by macroseeding. All drops contained 0.1M Tris buffer pH 8.4 and 8% PEG 8K as precipitant. The initial drop contained 3µl of protein at a concentration of 12 mg/ml and 3µl of 8% PEG 8K, 0.1M Tris pH 8.4. A 3µl volume of a small crystal suspension obtained from the initial drop were then dissolved in 100µl of the well solution and aspirated violently. 1µl of this solution was then added to 3µl of protein and 3µl of new well solution. Finally, the larger crystals obtained from this trial were washed in successive buffers containing increasing concentrations of PEG 8K up to 8% and macroseeded in a solution containing 1µl of protein and 3µl of well solution. Crystals up to 0.4 x 0.1 x 0.1mm were obtained after under a week of incubation at room temperature. Diffraction extended to 2.8Å from these crystals. However after approximately one month of incubation at room temperature, a crystal of superior quality was obtained from a hanging drop initially containing 3µl of 4% PEG 8K, 0.1M TRIS pH 8.2 and 3µl of protein of concentration 12mg/ml (fig. 2.3). The crystal dimensions were 0.2 x 0.5 x 0.5mm. Diffraction extended up to 2.5Å at the synchrotron from this crystal.

Fig 2.3 Higher quality 68kDa crystal.



Discussion

The purpose of purification in both cases was to provide protein of a suitably high purity to begin crystallization trials. From inspection of the SDS-PAGE gels, it can be seen that the 41kDa sample obtained was significantly purer than the 68kDa sample. Although it has been well documented that impurity can be a major reason behind the failure of proteins to crystallize (Abergel *et al.*, 1991), the purified 68kDa and 41kDa protein samples were of high enough purity to produce good quality crystals. The difficulty in obtaining good quality crystals of the 68kDa form of the enzyme could be attributed to impurities in the protein sample and/or ease of cleavage of the 68kDa protein to the 41kDa form. In fact, 41kDa crystals were repeatedly produced in 68kDa crystallization trials, sometimes in the same hanging

drop as 68kDa crystals. Affinity purification of sialidases has been reported (Suzuki *et al.*, 1993; Milligan *et al* 1980) and may be a technique that could be utilized in the future to improve purification of the 68kDa form of *M.viridifaciens* neuraminidase. Unfortunately, due to time constraints in the project, this method could not be developed in the purification reported here.

Chapter 3

NATIVE DATA COLLECTION AND ANALYSIS

Introduction

Once suitable crystals have been obtained, data collection can then begin. All data for the 41kDa crystals were obtained in-house on a Siemens Area Detector, whereas data for the 68kDa crystals were also collected using a synchrotron source at station X31, DESY, Hamburg, Germany on a 18cm MAR image plate. The XDS (Kabsch, 1988) suite of software was used for subsequent data analysis in all cases, except the final 68kDa data set where DENZO (Otwinowski, 1993) was used.

The Area detector

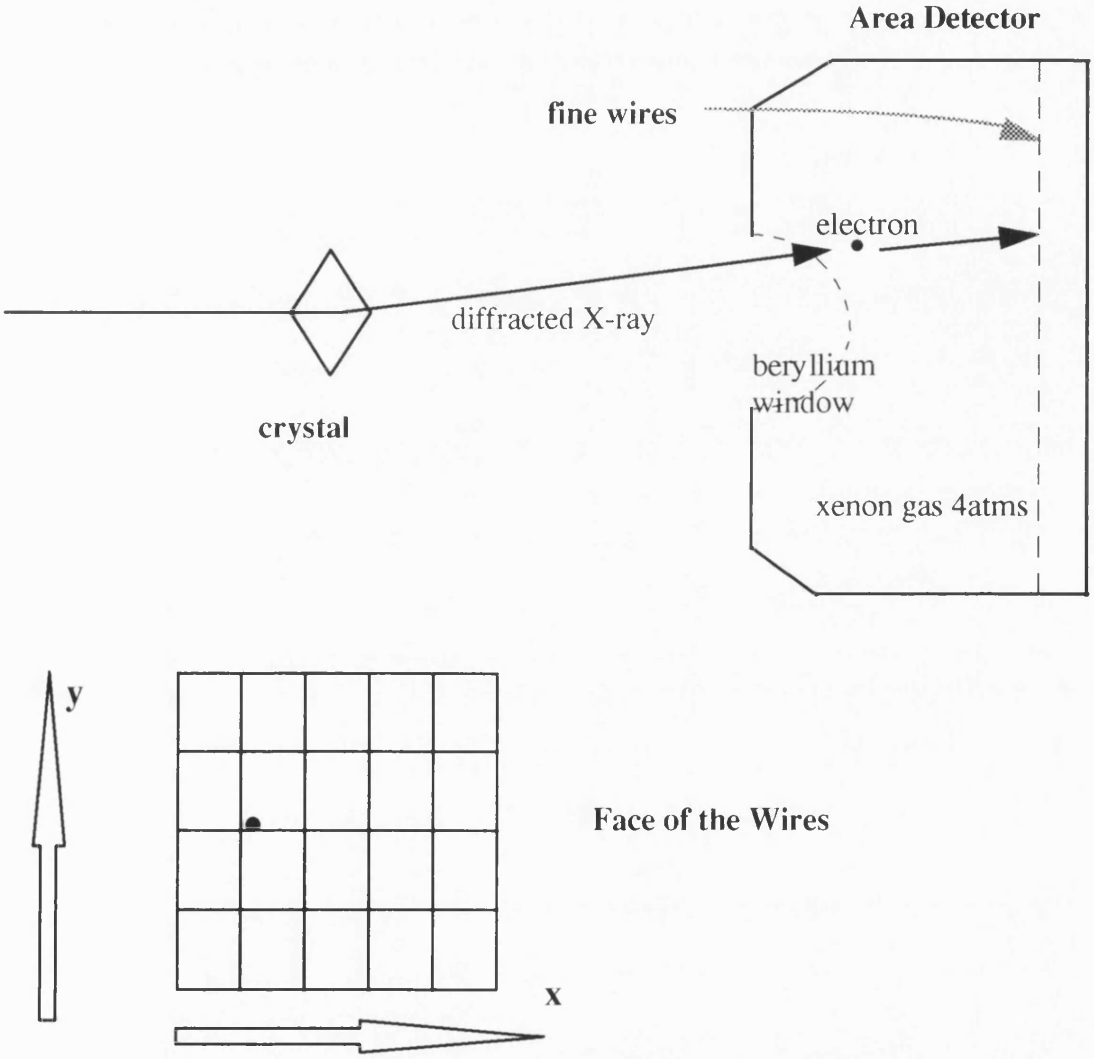
X-rays of wavelength 1.54Å are produced by a rotating copper anode X-ray source, with a graphite monochromator used to select Cu Kα radiation, typically operated at 35-40kV, 70-80mA. The detector type used in-house was a Siemens multiwire area detector (model:X-100). The detector is mounted on a moveable 2θ arm whose axis is coincident with the rotation ω axis of a 3-circle goniostat. A Motorola 68010 - based

PCS computer, running an UNIX operating system controls the goniostat and detector via the camera controller and position decoding microprocessors. A colour monitor, connected to the PCS, enables inspection of data as they are collected. The PCS is connected to a MicroVAX computer by an Ethernet link.

A schematic diagram of the area detector is shown in Fig. 3.1. X-ray photons cause ionization of xenon gas (at 4 atmospheres) in the detector chamber. The high pressure of the gas ensures that ionization events occur just behind the 1mm concave beryllium window, inhibiting spot spreading due to parallax. The concave nature of the beryllium window also helps to minimize parallax by ensuring the electric field lines are approximately parallel to the incident X-ray beams. Electrons produced by the ionization events are accelerated towards the anode wires, at approximately 5kV above ground, causing secondary ionization events and amplifying the signal. The cathodes consist of two parallel planes of wires - one with the wires running vertical and the other horizontal, enabling the determination of the position of the ionization events in two dimensions. The signals produced by the xenon cations as they reach the cathode wires are passed to the position decoder which decodes the signals as the (x, y) coordinate in the anode planes with an effective resolution of 0.2mm. The data are stored as a 512 x 512 pixel array (each pixel 0.2mm on an edge) in memory and written to disk at the end of the exposure. Efficiency of data collection depends on the count rate in two ways: (1) the decoder can store a second event that arrives whilst it is decoding the first, but loses the third and subsequent events; (2) the occurrence of high photon flux causes spots to spread out, reducing the effective resolution of the detector. The "% late" parameter reported by the PCS is a direct measure of the first

effect - it represents the number of events arriving at the PCS whilst the previous one is being decoded, as a proportion of the total counts recorded. The "% late" is also a monitor of the second effect since it is also a function of photon flux.

Fig. 3.1 Schematic diagram of the area detector. The intensity of the diffracted beam is determined from the number of electrons at each wire and the x and y value is determined for that position.



To eliminate local heterogeneity in the detector response due to irregularities in the wire spacing, a correction is performed using a radioactive ^{55}Fe X-ray source. A "flood field" correction is recorded with the detector being exposed to the ^{55}Fe X-radiation. The counts from each pixel are locally re-binned to give a smooth response over the detector surface using the program CORRECT. This correction does not compensate for any non-local sensitivity differences in the detector response. The correction is applied on-line whilst data are collected and thus recorded frames have already been corrected in this manner. An additional image is required before data collection - the "brass plate" image is collected using the ^{55}Fe source: a brass shield which has holes drilled at precise intervals is placed over the detector face. The distribution of the spots in the brass plate image in comparison with their known position on the brass plate enables a correction for geometric distortion to be made.

Data are collected as a series of consecutive non-overlapping "frames", with the crystal oscillating about the vertical ω axis over a small angular range. Each diffraction point is divided into a number of slices in ω according to the angular width of a frame ($\Delta\omega$), from which the three-dimensional profile of the reflection can be reconstructed. The small angular range also increases the signal-to-noise ratio of the reflections. The crystal is oscillated about the ω axis, rather than simply rotated, to minimize errors due to fluctuations in the X-ray beam and because the opening of the shutter and the ω axis motors are not synchronized.

Each frame is stored as a 512x512 byte array with appropriate header information on exposure time etc. Pixels which contain more than 255 counts have their

true count stored in the header and are flagged as "overflows". The number of overflows on a frame thus depends on the number of diffraction maxima and their strength. Frames are transferred from memory directly to the MicroVAX disk.

Data collection in house

Each crystal was drawn into a thin-walled glass capillary tube, with a minimum amount of mother liquor from the hanging drop and the tube sealed with beeswax. The capillary tube was then mounted on the goniometer head and secured with plasticine. The crystal was positioned in the X-ray beam by adjusting the translational sledges whilst observing the crystal through a telescope.

The minimum crystal to detector distance in centimetres required to be able to resolve individual spots is roughly the length of the largest cell axis (in Å) divided by 10 (G.L. Taylor-personal communication). Initially, the crystal to detector distance should be relatively large in order to ensure that individual spots do not overlap. However, after determination of the unit cell dimensions from data processing, the distance can be reduced so that higher resolution data can be collected. The 2θ swing angle in the goniometer coordinate system and the crystal to detector distance determine the maximum resolution of data that can be achieved.

$$d_{max} = \frac{\lambda}{2\sin\theta_{max}}$$

$$2\theta_{max} = \theta_c + \tan^{-1}\left(\frac{r}{D}\right)$$

d_{max} = maximum resolution, D=crystal to detector distance, r = radius of detector,

θ_c = detector swing, θ_{max} = angle of maximum resolution, λ = wavelength of X-ray beam.

All data were collected in a series of contiguous, non-overlapping frames, with each frame representing an oscillation around ω of 0.25°. All data were collected at room temperature and reduced using the XDS suite of software (Kabsch, 1988).

XDS data reduction

The XDS data processing program accepts frames of data from the Xentronics area detector / MAR image plate and outputs a list of squared structure factor amplitudes, together with information about the completeness of data, quality and optionally agreement with a reference data set. The program consists of initialization, reflection profile collection and post-processing steps. The specific steps are described below and summarized in the flow-diagram (fig. 3.2).

1. XYCORR reads the latest brass plate image and determines the radial distortion, as described above, thus making an allowance for the difference between the observed and known spot positions. A look-up table is created, which is used for correcting for radial distortion in each image.

2. INIT reads the first 30 frames of data and calculates the average background at each detector pixel. For a given pixel position, each reading is accepted if it is less than three standard deviations above the lowest of the readings. The average of these accepted readings is then taken as background, so that strong diffraction spots are not erroneously taken as background.

3. COLSPOT locates the centroids of strong spots which are significantly over background, for a given number of images. A list of observed spot and spindle positions is produced for autoindexing.

4. IDXREF tries to find the unit cell dimensions of the crystal and its orientation with respect to the X-ray beam and detector, by reading the list of spot coordinates and spindle positions produced by Colspot. The direct beam position and unit cell dimensions are refined.

5. COLPROF collects reflection profiles from all the data frames. Crystal orientation and unit cell dimensions are refined every 15 frames, using all reflections collected in the previous 15 frames. A 3-dimensional reflection profile is generated, represented in the output by a simplified box of 9 x 9 x 9 grid points. A good data set typically has a compact reflection profile whose centroid is close to the centre of the box.

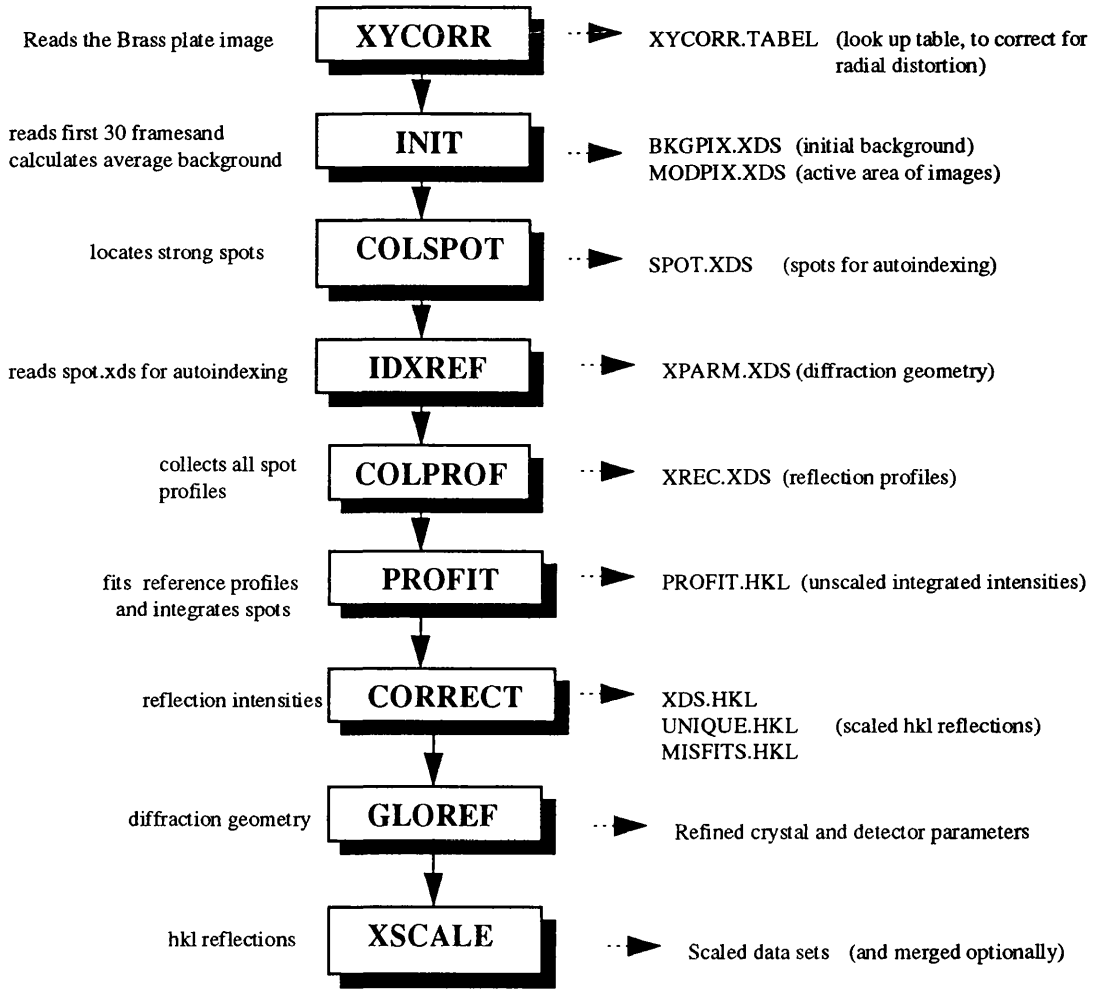
6. PROFIT performs profile fitting. Signal and background points are distinguished, assuming that weak and strong reflections have the same normalized profile. Reflection intensities are estimated by a two pass integration procedure. In the first pass, reference profiles are learned as a function of position on the detector and spindle angle within the data collection range. In the second pass, intensities are estimated for each reflection using the nearest reference profile.

7. CORRECT scales the data to correct for X-ray damage to the crystal and variation in detector sensitivity. Correct outputs scaled intensities for reflections and their standard deviations.

8. GLOREF performs post refinement of all instrument parameters and unit cell parameters using all strong reflections.

9. XSCALE can scale and merge multiple data sets and can apply a semi-empirical absorption correction.

Fig. 3.2 Flow diagram summarizing data analysis using XDS.



Synchrotron data collection

A typical modern synchrotron radiation source consists of a storage ring in which charged particles (electrons or positrons) are circulated for long periods of time. These particles are accelerated by a linear accelerator and then injected either directly into the

storage ring where they are accelerated to a desired energy or into a booster storage ring where the energy is increased before injection into the storage ring. In the storage ring the particles travel through bending magnets that define their path and straight sections where the bunch shape is defined by quadrupole or sextupole magnets. The intensity of the synchroton radiation is increased by Wigglers, which are a series of magnets that cause the particle beam to undergo multiple transverse excursions from the normal pathway. As a result, the energy of the radiation is increased and the spectrum is shifted to smaller wavelengths. The use of synchrotron radiation as opposed to an in-house source gives several distinct advantages:

1. Absorption and radiation damage are reduced, due to the shorter wavelengths of synchrotron radiation which can be used.
2. The wavelength of the synchrotron radiation can be tuned, due to the broad spectrum that is emitted. Applications that can exploit this are Laue crystallography (Pai, 1992), and multiple wavelength anomalous diffraction (MAD) phasing (Hendrickson, 1991).
3. Higher resolution data can be collected due to the high intensity of the X-rays. In addition, since the radiation is more intense, a data set can be collected to a high degree of completeness more quickly.

Data processing was again performed using the XDS suite of software as described above, but the final 68kDa data set was reduced using DENZO. **DENZO** performs profile fitting to all the predicted spots on the images collected. The user can determine the size of the box, in which the spot and background are integrated. A χ^2 test is calculated for every reflection, where an overall value of less than 2 indicates

good data processing. DENZO can also refine all the crystal and instrument parameters simultaneously. These include unit cell dimensions, wavelength used, direct beam coordinates and crystal to detector distance. **SCALE** then merges all the processed images, applying the appropriate correction factors and determines the mosaicity of the crystal.

Data analysis

41kDa Crystals

Parameters for data collection were as follows:

crystal to detector distance = 12cm

X-ray generator settings = 50kV, 80mA

exposure time = 120 - 180 seconds/frame

2θ swing angle = 30°

temperature ≈ 21°C

The source of protein for all the 41kDa crystals from which data were collected was the culture filtrate.

Diffraction extended to 1.82Å from the in-house rotating anode source for native crystals of the 41kDa form of the enzyme and a summary of the data collection statistics are shown in table 3.1. This data file consisted of five merged data sets, each obtained from one crystal of dimension approximately 1.6 mm x 0.6 mm x 0.6 mm. The crystals were judged to be of space group $P2_12_12_1$, from observation of systematically absent or weak axial reflections.

The unit cell volume = 337527Å^3, molecular mass = 41kDa/monomer. So if we assume that there is one monomer in the asymmetric unit:

$V_m = \frac{337527}{41000 \times 4} = 2.03$Å^3/Da which is in the reported range given by Matthews for crystals (Matthews, 1968). Since the fractional volume occupied by solvent in the crystal = $1-\frac{1.23}{Vm}$, the solvent content of the crystal was estimated to be 38.5%.

Table 3.1 Summary of data collection statistics for the 41kDa crystal.

resolution (Å)	no. of observations	no. of unique reflections	completeness >1σ (%)	completeness >3σ (%)	R_I (%)[a]
∞ - 9.0	5	4	1.3	1.3	3.3
9.0 - 6.0	3762	658	98.5	95.2	4.6
6.0 - 4.0	15999	2148	100.0	98.3	4.7
4.0 - 3.0	30907	4082	99.3	97.4	5.1
3.0 - 2.0	95748	16363	98.3	92.0	6.3
2.0 - 1.82	21875	6566	81.1	63.9	9.4
∞ - 1.82	168,296	29,821	93.4	85.5	5.5

cell dimensions	a	b	c	α	β	γ
	48.14	82.73	84.75	90.0	90.0	90.0

$$^aR_I = \frac{\sum_h \sum_i |(I_h - I_{hi})|}{\sum_h \sum_i I_{hi}}$$

where I_h is the weighted mean measured intensity of the observations I_{hi} in which the intensities of the symmetry related reflections, which should be the same, are compared. R_I, or the merging R factor gives an estimate of their disagreement.

68kDa Crystals

The native data set was collected for the 68kDa crystals from: 1. In-house 4Å data set (September 1994) 2. Hamburg synchrotron 2.8Å data set (November 1994) and 3. Hamburg 2.5Å data set (May 1995).

Parameters for data collected in house were as follows:

crystal to detector distance = 12cm

X-ray generator settings = 40kV, 100mA

exposure time = 5minutes/frame

2θ swing angle = 5°

temperature ≈ 21°C

The source of protein for all the 68kDa crystals from which data were collected was the culture filtrate.

Initially, crystals of the 68kDa form of dimension approximately 0.4mm x 0.1mm x 0.1mm only diffracted to 4Å on the in-house source (table 3.2). However, diffraction from crystals of the same size extended to 2.8Å at the DESY synchrotron, Hamburg, on station X31 at a wavelength of 0.993Å (table 3.3). Due to time constraints though, only 54° of data were collected in 1° frames from one crystal, with exposure times of between 170 and 300 seconds. After merging of the in-house and synchrotron data sets with XSCALE , a data set was obtained (table 3.4) with a completeness of 57.2% at 3.5Å (>1σ).

A larger 68kDa crystal of dimensions 0.5 x 0.4 x 0.2 mm was used at a later date, to collect a further Hamburg data set, which was 80.7% complete to 2.5Å (>1σ). Data processing statistics are shown in Table 3.5.

After data analysis using XDS, the crystals of the 68kDa enzyme were determined to be of space group $P2_1$, as judged by systematic absences of the type 0k0: k = 2n.

The unit cell volume = 349650Å^3, molecular mass = 68kDa/monomer. So if we assume that there is one monomer in the asymmetric unit:

68kDa occupies a volume of 349650Å^3, so $\frac{349650}{68 \times 2} = 2.57$Å^3/Da which is in the reported range given by Matthews for crystals. Since the volume occupied by solvent in the crystal = $1-\frac{1.23}{Vm}$, the solvent content of the crystal was estimated to be 52%.

Table 3.2 In-house 68kDa native data set processing statistics.

resolution (Å)	no. of observations	no. of unique reflections	completeness (>1σ) %	completeness (>3σ) %	R_I (%)[a]
∞ - 10.0	740	361	89.6	87.3	2.7
10.0 - 6.0	2972	1377	93.5	80.6	6.0
6.0 - 5.5	1068	522	91.7	73.8	8.4
5.5 - 5.0	1172	745	92.5	73.9	7.2
5.0 - 4.5	1301	1084	88.9	70.4	4.3
4.5 - 4.0	1723	1488	76.9	50.9	5.6
∞ - 4.0	8976	5577	87.2	68.9	5.5

cell dimensions	a	b	c	α	β	γ
	51.13	117.10	60.01	90.0	95.58	90.0

Table 3.3 Hamburg (November 1994) synchrotron 68kDa native data set processing statistics.

resolution (Å)	no. of observations	no. of unique reflections	completenes s (>1σ) %	completeness (>3σ) %	R_I (%)[a]
∞ - 10.0	1048	362	89.1	87.4	4.8
10.0 - 6.0	4403	1384	93.2	82.7	8.0
6.0 - 5.0	3562	1273	92.9	79.7	10.0
5.0 - 4.0	5989	2652	85.7	70.0	9.8
4.0 - 3.5	4351	1870	57.2	40.1	12.7
3.5 - 2.8	8443	3501	31.3	18.9	17.7
∞ - 2.8	27,796	11,042	55.7	42.3	10.3

cell dimensions	a	b	c	α	β	γ
	51.03	117.41	60.07	90.00	95.62	90.00

Table 3.4 Merged Hamburg (November 1994) synchrotron and in-house 68kDa native data set processing statistics

resolution (Å)	no. of observations	no. of unique reflections	completeness >1σ (%)	completeness >3σ (%)	R_I (%)[a]
∞ - 10	1,048	362	89.1	87.4	4.8
10 - 6	4,403	1,384	93.2	82.7	8.0
6 - 4.5	6,010	2,357	91.7	78.9	9.2
4.5 - 4.0	3,541	1,568	82.6	64.8	10.9
4.0 - 3.5	4,351	1,870	57.2	40.1	12.7
3.5 - 3.0	5,475	2,210	34.3	22.9	15.4
∞ - 3.0	24,828	9,751	62.4	49.1	9.9

cell dimensions	a	b	c	α	β	γ
	51.03	117.41	60.07	90.00	95.62	90.00

Table 3.5 Hamburg (May 1995) synchrotron 68kDa native data set processing statistics. A total of 63,889 reflections were used.

resolution (Å)	no. of unique reflections	completeness (>1σ) %	completeness (>3σ) %	R_I (%)[a]
4.27 - 3.97	984	80.5	80.2	3.0
3.55 - 3.39	1004	81.6	81.3	3.7
3.15 - 3.05	1003	81.8	81.8	4.5
2.89 - 2.82	1024	80.8	79.9	5.9
2.69 - 2.64	1026	81.6	80.4	7.7
2.54 - 2.5	974	77.4	71.7	10.4
∞ - 2.5	20083	80.7	80.0	3.7

cell dimensions	a	b	c	α	β	γ
	51.28	117.30	60.24	90.0	96.17	90.0

$$^{a}R_I = \frac{\sum_h \sum_i |(I_h - I_{hi})|}{\sum_h \sum_i I_{hi}}$$

Discussion

A high quality native data set for the 41kDa form was obtained in-house. However, it was necessary to collect data at the synchrotron for the 68kDa crystal. As described above, due to time constraints the first synchrotron data set collected from the 68kDa crystal contained only 54° degrees of data and this is reflected in the completeness of the final merged data set. After the initial difficulty in obtaining higher quality crystals, a 2.5Å 68kDa data set was finally obtained and this was used for solving the 68kDa form by molecular replacement. Cryogenic techniques were also attempted in-house, in order to extend the resolution limit of the 68kDa crystals, but were unsuccessful.

Chapter 4

ISOMORPHOUS REPLACEMENT

Introduction

For macromolecular structure determination an interpretable electron density map is essential, into which the polypeptide backbone of the protein can be built, and the side chains inserted. The calculation of such a map requires that both the amplitudes and the phases of the diffracted X-rays are known since the electron density equation is given by the formula:

$$\rho(x,y,z) = \frac{1}{V} \sum_{hkl} F_{hkl}\, e^{i\alpha_{hkl}}\, e^{-2\pi i(hx+ky+lz)}$$

The amplitudes of the diffracted rays can be obtained from the intensities of the spots in the diffraction pattern since amplitude (F) is directly proportional to the square root of the intensity. However all the phase information is lost in the diffraction experiment, since there is no lens capable of focusing the X-rays. This is known as the phase problem and is solved indirectly in macromolecular crystallography via either isomorphous replacement or molecular replacement. Molecular replacement will be discussed in Chapter 6 and in Appendix 2.

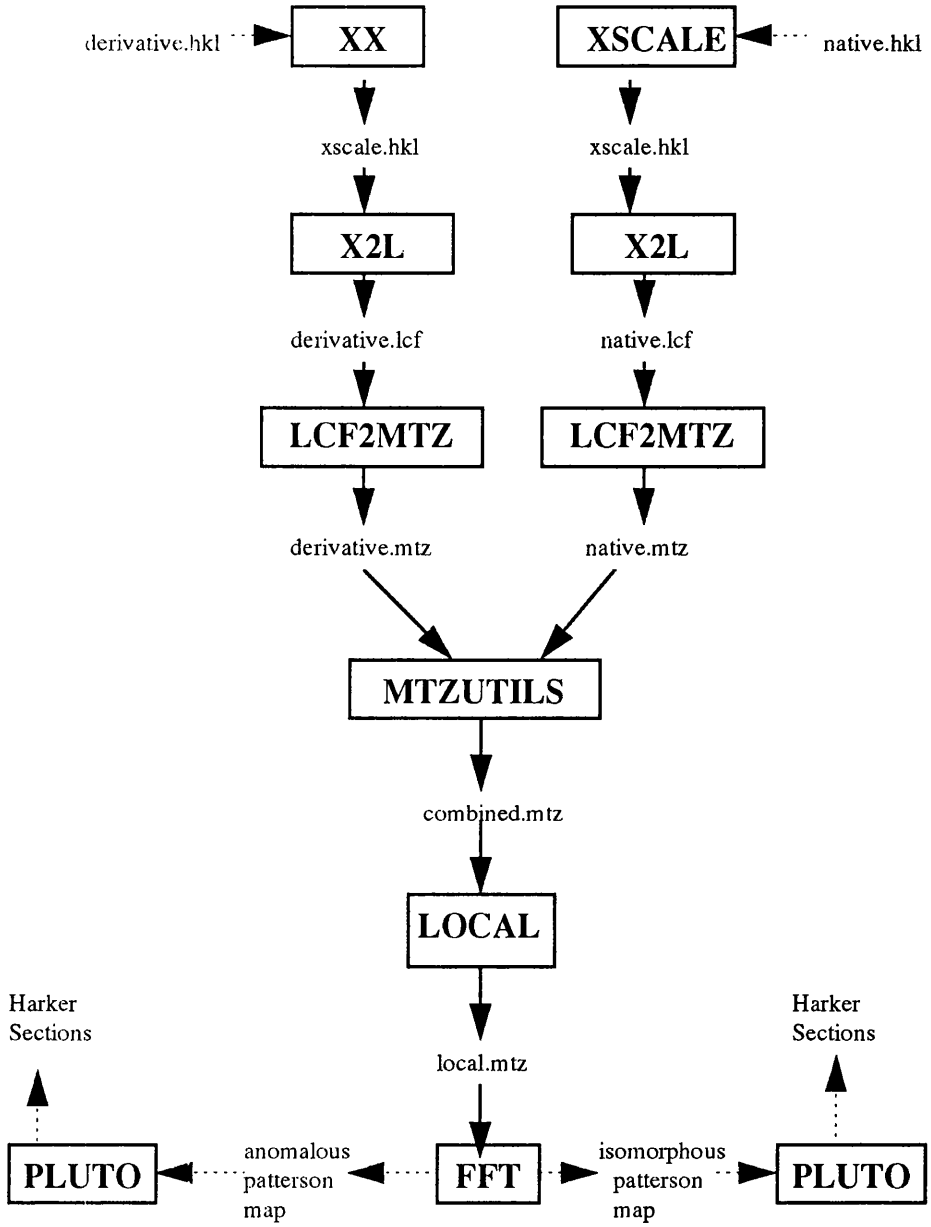
The technique of isomorphous replacement was employed in the structure elucidation of the 41kDa form of *M.viridifaciens* neuraminidase. There are generally

large solvent channels present within protein crystals offering routes along which heavy atoms can diffuse, to reach molecules throughout the crystal. In order to obtain a suitable derivative the heavy atoms should bind to the same site(s) on every molecule. The addition of a sufficiently electron dense heavy atom will produce observable changes in the diffraction pattern. The position of the heavy atoms can be located using the Patterson method and from the intensities of the native and derivative diffraction patterns, the phase problem can be solved (Appendix 2). For centrosymmetric crystals, a single heavy atom derivative will be sufficient to unambiguously determine the phase. However there is ambiguity in the phases for non-centrosymmetric crystals and this is generally overcome by using at least one further derivative. The search for heavy atom derivatives is generally by a "trial and error" process and hence can often be a time-consuming, arduous task. Hints about possible heavy atom derivatives can be obtained from the amino-acid sequence, e.g. binding of mercury compounds to sulphydryl groups, but without a detailed knowledge of the crystal structure, it is not possible to have a completely rational approach. A derivative may be obtained in one of two ways; either by soaking a solution of the heavy atoms into the crystal, or by cocrystallizing in the presence of heavy atoms.

Obtaining phases from heavy atom derivatives

The derivative hkl reflection file obtained in the way described in Chapter 3, is merged to a unique set of data using a modified version of XSCALE, called XX, which preserves the Friedel pair intensities. Figure 4.1 shows the steps employed in obtaining Patterson maps to locate the heavy atom positions.

Fig. 4.1 Flow diagram illustrating the steps involved in generating Pattersons.



The derivative and native hkl reflection files are converted to lcf format using X2L, which can then be converted to mtz format with LCF2MTZ for use in the CCP4 suite of programs. The mtz file contains structure factor amplitudes and their standard deviations and anomalous differences and their standard deviations, if present for the

derivative, for all reflections of the type hkl. The derivative and native mtz reflection files are merged in MTZUTILS and the derivative is scaled to the native using LOCAL.

LOCAL scales the combined data sets in two distinct stages. Firstly, an overall scaling is performed which uses an anisotropic temperature factor followed by local scaling. Thus the data are reasonably well scaled before local scaling is employed. LOCAL can also scale Friedel pair differences, arising from anomalous scattering by heavy atoms in the derivative. The scaled mtz file is read into FFT which outputs an isomorphous difference Patterson map. The Patterson map is generated by the synthesis:

$$P\ (u,v,w) = \frac{1}{V} \sum_{hkl} (F_{PH}-F_P)^2\, e^{-2\pi i(hu+kv+lw)}$$

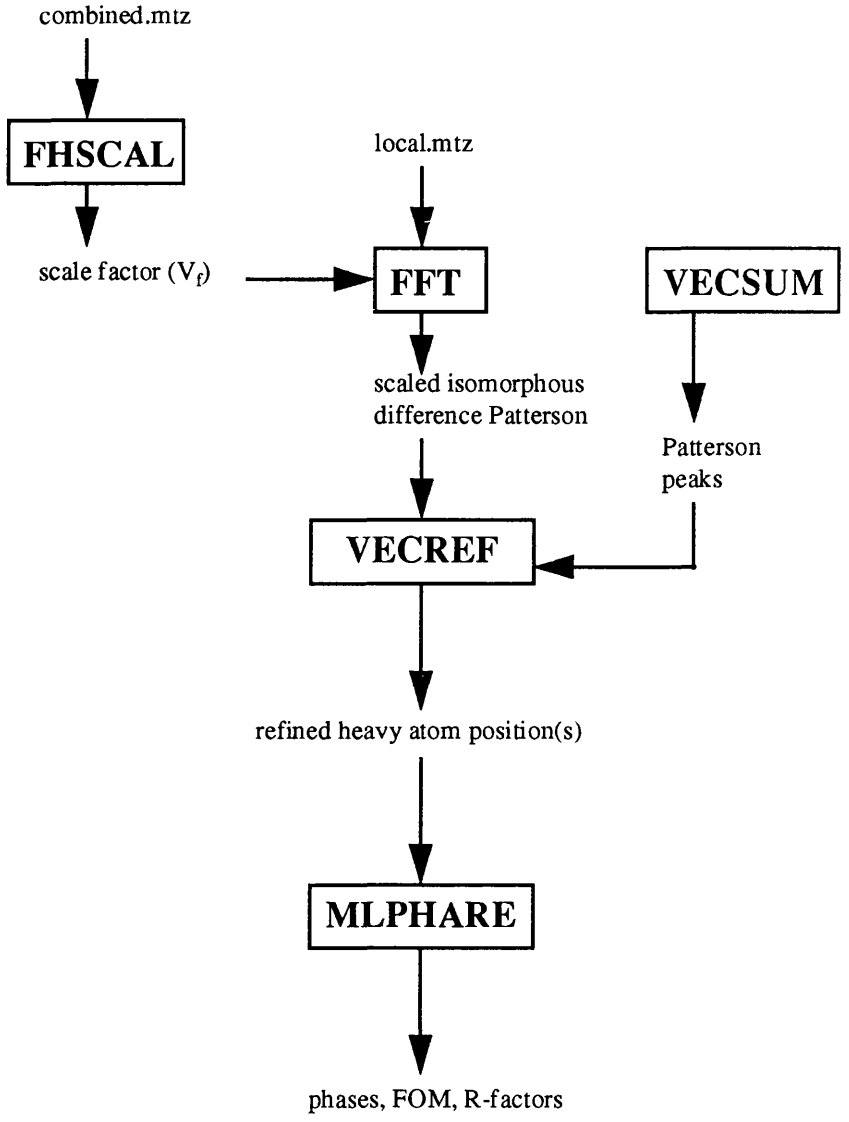
FFT can also generate the anomalous difference Patterson map, calculated with respect to disparity between the Bijovet pairs. The Patterson synthesis here is the same as for the above, except with coefficients $(F_{PH}(+) - F_{PH}(-))^2$.

PLUTO plots sections of the Patterson maps. In the case of the 41kDa form, the space group was $P2_12_12_1$ and Harker sections lie at u=1/2, v=1/2, w=1/2.

The isomorphous difference Patterson or anomalous difference Patterson map produced by the FFT is then read into VECSUM, which deconvolutes the Patterson function and produces a map of the possible heavy atom sites in the structure. Once an initial possible heavy atom site has been located, its coordinates are put back into VECSUM and further sites searched for. All genuine peaks were manually located on the Harker sections. A summary of the steps involved in location and refinement of heavy atom sites in order to obtain phase information is shown in Fig. 4.2.

The merged native and derivative (combined.mtz) reflection file from MTZUTILS is read by FHSCAL which scales the data using Kraut's formula (Kraut *et al.*, 1962) and outputs a scale factor (V_f). This is used in FFT to produce a correctly scaled isomorphous difference Patterson. VECREF then refines the putative heavy atom positions, reading the scaled isomorphous difference Patterson map and the putative heavy atom site coordinates. MLPHARE refines the heavy atom positions and calculates phases, giving a variety of statistics regarding the quality of the phase information.

Fig. 4.2 Location and refinement of heavy atom sites, in order to obtain phases.



The Search for successful derivatives

Crystals were soaked in solutions of Gold, Platinum, Iodine, Thorium, Cadmium and Lanthanum. Table 4.1 lists the unit cells for the derivatised crystals.

Table 4.1 Summary of some of the unit cells of the derivatised crystals.

	a (Å)	b (Å)	c (Å)	α=β=γ (°)
native	48.14	82.73	84.75	90
K_2PtCl_6	48.00	82.97	85.35	90
$Th(NO_3)_2$	48.11	83.14	85.06	90
$NaAuCl_4$	48.05	82.95	85.23	90
$CdCl_2$	48.02	83.10	85.01	90
KI_3	48.11	82.92	85.12	90
$La(CH_3COO)_3$	48.15	83.16	85.07	90

a) Unsuccessful derivatives

Heavy atom soaks which resulted in either non-isomorphism or a native crystal are listed briefly here. All possible peaks were verified in VECREF, but the derivative was not pursued when the occupancies refined to zero and gave poor correlations and residuals.

1. K_2PtCl_6 - soaked for 2 days at a concentration of 2mM. Possible sites from VECSUM gave poor correlations and residuals in VECREF. The crystal was therefore deemed to be native.

2. $NaAuCl_4$ - soaked for 2days and for 1 day at a concentration of 2mM. Both conditions resulted in poor correlations and residuals in VECREF. Both crystals were therefore not derivatives.

3. **$Th(NO_3)_2$** - soaked for 1.5 hours at a concentration of 2mM. The data set for this soak resulted in a fractional intensity difference of only 5.3%. In addition, the isomorphous difference Pattersons showed no peaks.

4. **KI_3** - soaked for 1.5 hours at a concentration of 5mM. The Pattersons obtained from this data set showed no obvious peaks. The map was "smeared out" indicating possible non-isomorphism.

b) Successful derivatives

Data collection statistics for the derivatives which gave useful phase information are summarised in Table 4.2. Refinement statistics for the heavy atom sites are summarised in Table 4.3. The grid used in VECSUM was 48 44 44 for a resolution of 3.0Å.

1. $CdCl_2$ - soaked for 2 days at a concentration of 2mM. Two potential sites from VECSUM with grid coordinates 26 8 25 and 2 35 19 were found. After refinement in VECREF one major and one minor site was identified respectively. The cross vector peak corresponding to the two sites was also identified.

2. **$La(CH_3COO)_3$** - soaked for 3 days at a concentration of 2mM. One potential heavy atom site was identified from VECSUM with grid coordinates 21 17 13 and refinement in VECREF gave reasonable correlations and residuals indicating one major heavy atom site.

3. KI_3 - soaked for 10 days at a concentration of 0.2mM.

(as for the Lanthanum derivative, with VECSUM grid coordinates 18 19 35)

4. $Th(NO_3)_2$ - soaked for 2 days at a concentration of 2mM. One major peak was identified in VECSUM with grid coordinates 24 18 15 and subsequent refinement in VECREF gave reasonable residuals and correlations as well as a high occupancy, reflecting the peak height in the isomorphous difference Patterson (7σ in w=1/2 Harker section).

Table 4.2 Data collection statistics for the heavy atom derivatives.

derivative	$Th(NO_3)_2$	$CdCl_2$	KI_3	$La(ac)_3$
No. of observations	24330	24240	24577	24521
No.of unique reflections	6589	6327	6814	6665
Completeness (%)	93.8	90.8	98.7	92.0
Resolution (Å)	3.0	3.0	3.0	3.0
R_I[a] (%)	6.9	6.9	10.4	10.5

[a] $$R_I = \frac{\sum_h \sum_i |(I_h - I_{hi})|}{\sum_h \sum_i I_{hi}}$$

Table 4.3 Summary of statistics for the heavy atom sites, together with the concomitant changes in cell dimensions (%Δ).

derivative site	rms fractional intensity difference (LOCAL)	fractional coordinates x y z	residual (VECREF)	correlation (VECREF)	relative occupancy (VECREF)	max peak height*(σ)	%Δ a b c
Cd 1		0.54, 0.09, 0.28	0.431	0.862	1.35	8.0	
Cd 2	0.197	0.04, 0.40, 0.20	0.960	0.187	0.22	1.0	0.24 0.45 0.31
I	0.148	0.36, 0.22, 0.40	0.611	0.666	0.33	3.0	0.06 0.23 0.43
La	0.260	0.46, 0.20, 0.15	0.560	0.681	0.38	5.0	0.02 0.51 0.38
Th	0.473	0.50, 0.20, 0.17	0.656	0.707	3.20	5.0	0.06 0.49 0.37

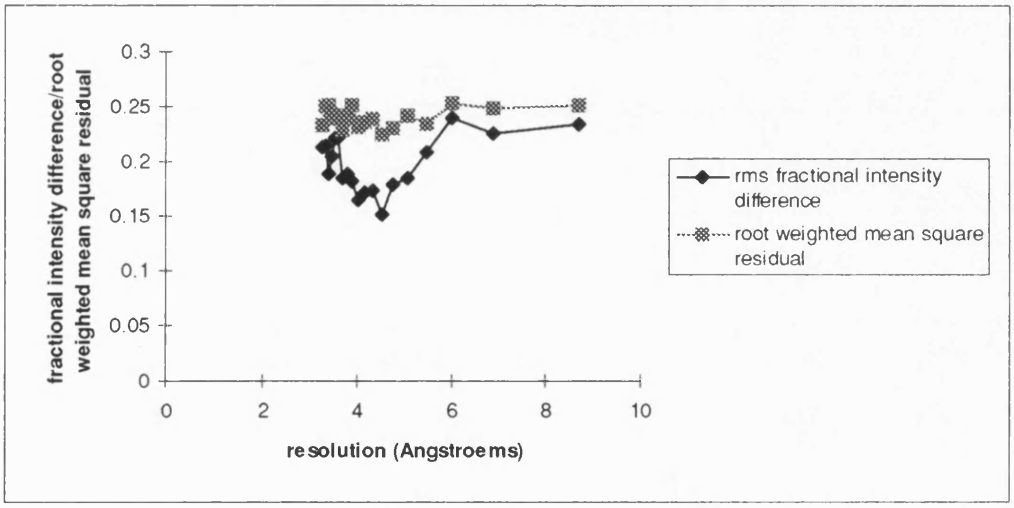
* max peak height is defined here by the maximum sigma level on the Harker section u=1/2

The scaling statistics for the derivatives of Cadmium, Thorium, Iodine and Lanthanum are illustrated graphically in Fig. 4.3.

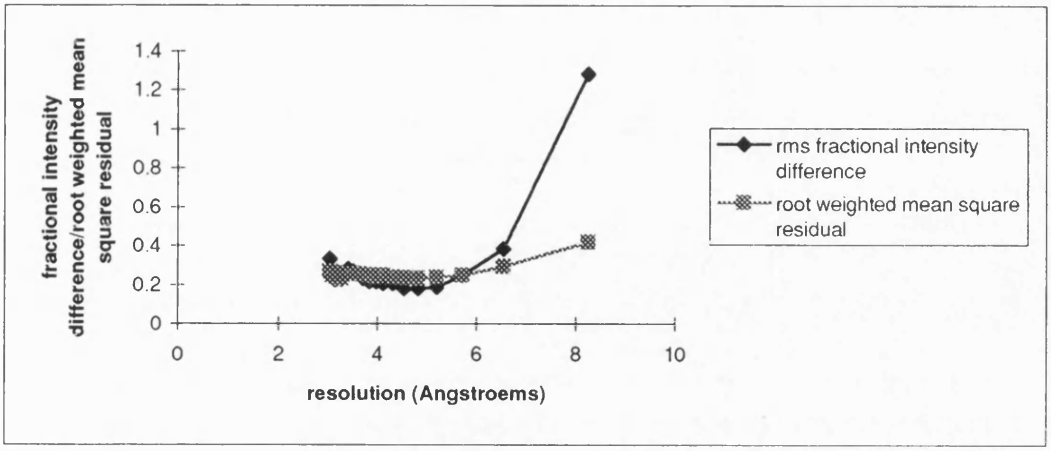
The isomorphous difference Pattersons for the section u=1/2 are illustrated in Fig. 4.4, for the successful Lanthanum, Cadmium, Thorium and Iodine derivatives. The maps are all contoured at 1σ and peaks due to heavy atom sites are labelled as is the cross vector peak for the Cadmium derivative.

Fig. 4.3 Scaling statistics for the successful derivatives. Rms fractional intensity difference and root weighted mean square residual are plotted as a function of resolution.

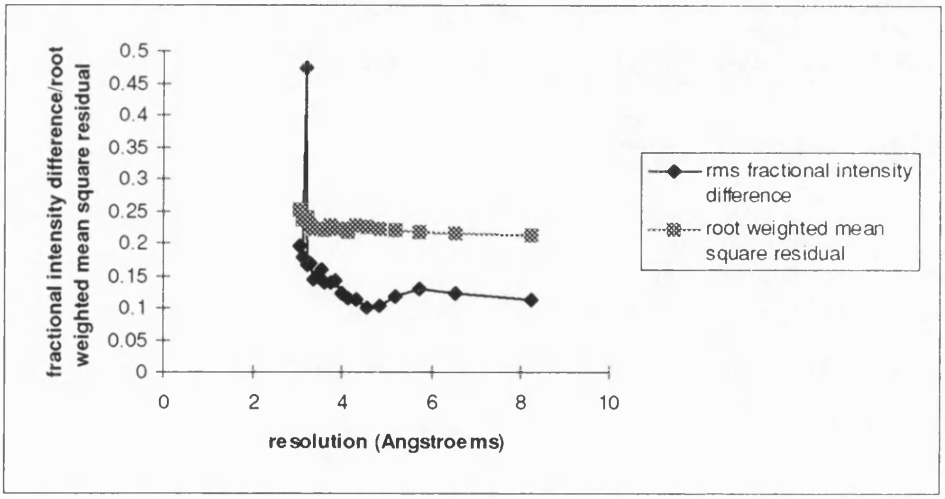
(a) Cadmium derivative



(b)Thorium derivative



(c)Iodine derivative



(d)Lanthanum derivative

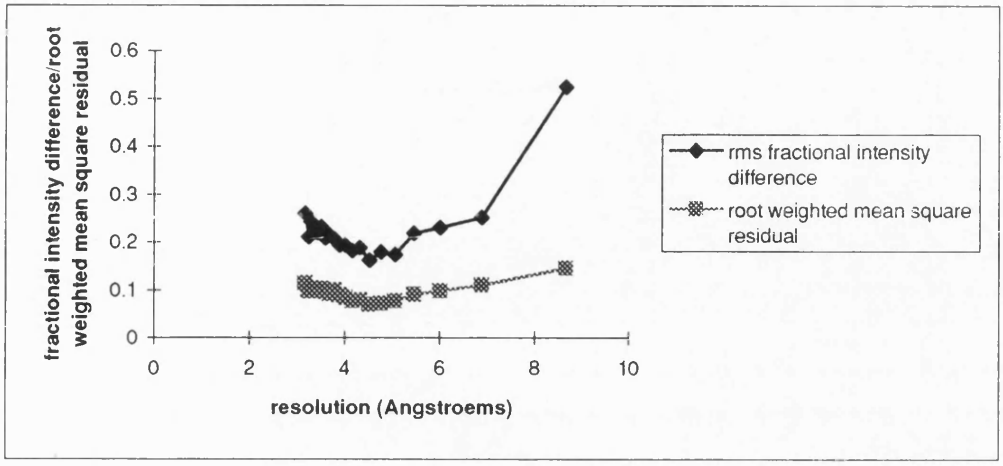
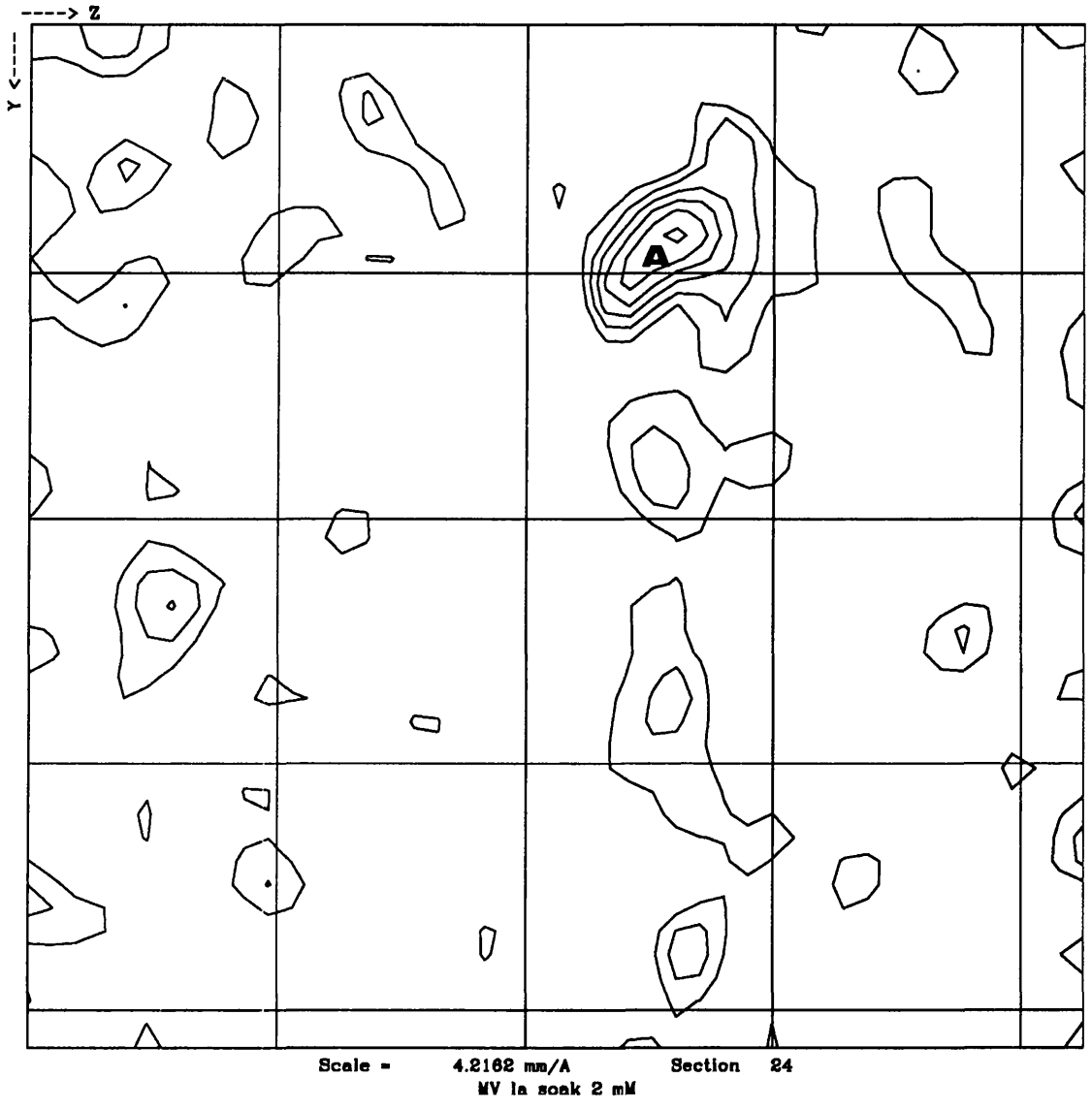
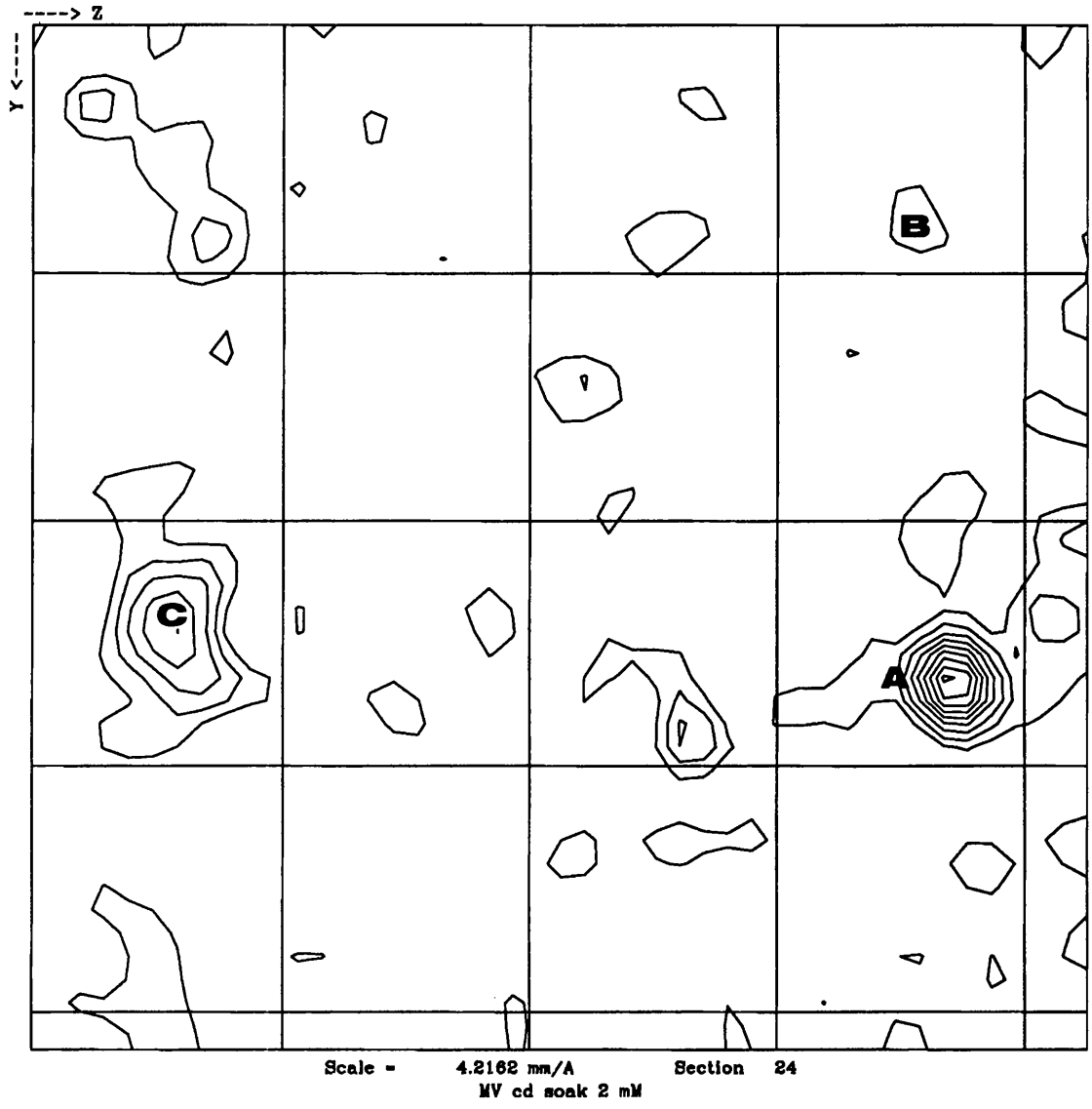


Fig. 4.4 u=1/2 Harker sections at 3Å for the successful derivatives. For each section A = peak corresponding to the heavy atom site. For the Cadmium Harker section, A = major site, B = minor site and C = cross vector peak.

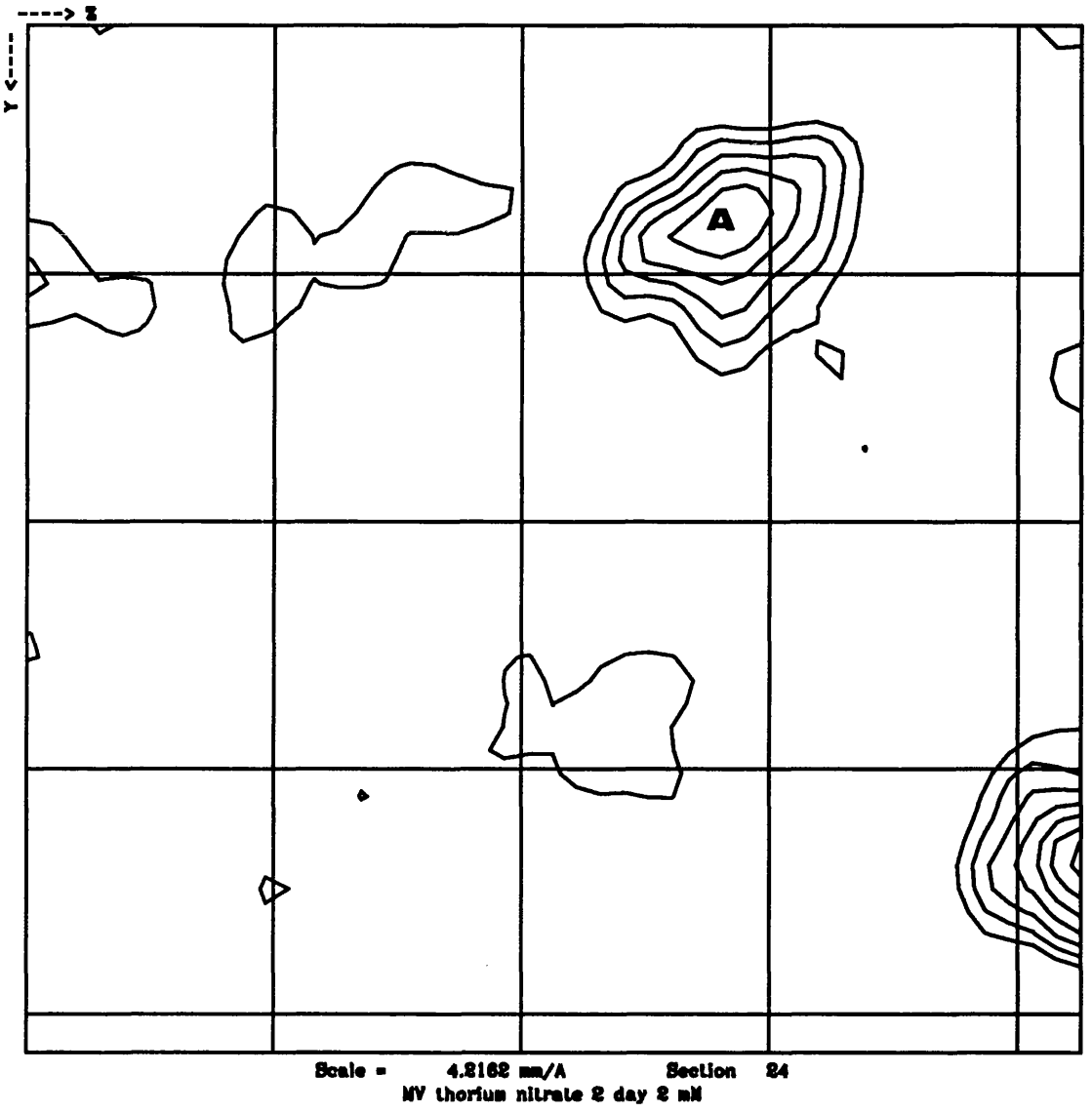
(a) Lanthanum derivative



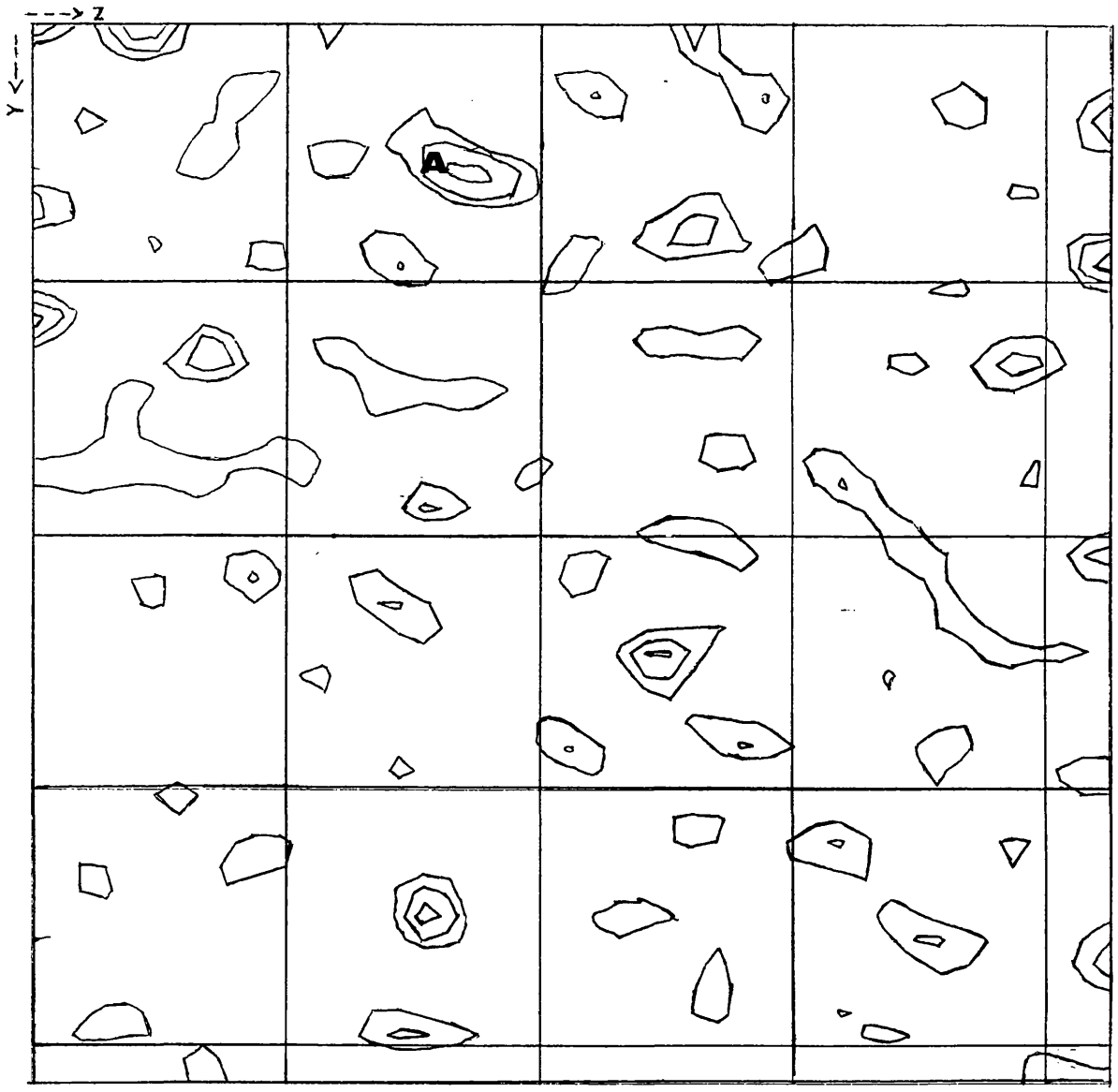
(b)Cadmium derivative



(c)Thorium derivative



(d)Iodine derivative



Scale = 4.2162 mm/A Section 24
MV ki3 soak 0.2 mM

Anomalous data

Anomalous difference Patterson maps were also calculated, but there were no significant peaks on the Harker sections u=1/2, v=1/2 and w=1/2 due to anomalous scattering. In addition, minor peaks on the anomalous Harker sections did not correspond to heavy atom peaks found on the isomorphous difference Pattersons. Hence, anomalous data were not used in future phase calculations.

Obtaining initial phase estimates

All phase estimates were calculated using MLPHARE. The Lanthanum derivative gave the first set of phases and then as more derivatives were found, their phases were calculated relative to those of the Lanthanum's. When a peak from VECSUM due to the presence of a heavy atom site is chosen, the origin and hand is arbitrarily fixed and there is only a 1 in eight chance that the peak due to a second heavy atom derivative will have the same origin. The first run of VECSUM shows all potential heavy atom sites, with respect to all origins and both hands. The Cadmium derivative gave two sites and so one Cadmium site was chosen and fed back into VECSUM. The origin and the hand was thus fixed and the second Cadmium site found was automatically on the same hand and origin as the first site.

Cross phase difference Fouriers with coefficients $\rho\ (x, y, z) = \sum_{hkl} FOM\ (F_{new\ deriv} - F_o)e^{iPHIB}$ were calculated for all the derivatives. Initially, PHIB corresponded to the best phases obtained from MLPHARE for the Lanthanum derivative, FOM corresponded to the figure of merits for the Lanthanum derivative and $F_{new\ deriv}$ was

the structure factor amplitude for the Iodine derivative, the second useful derivative found. When the next derivative was to be cross phased, PHIB corresponded to the best combined phases from the Lanthanum and the Iodine derivative, FOM corresponded to the figure of merit for the combined Lanthanum and Iodine derivative phases and $F_{new\ deriv}$ corresponded to that next derivative. Cross phase difference Fouriers were repeated in this fashion for the other derivatives that became available. Hence all heavy atom sites were put on the same hand and same origin as the Lanthanum's. Although the phases from the Thorium derivative were initially included, it was found that the overall phasing statistics were not significantly improved and therefore the Thorium derivative was omitted from the final phase calculations.

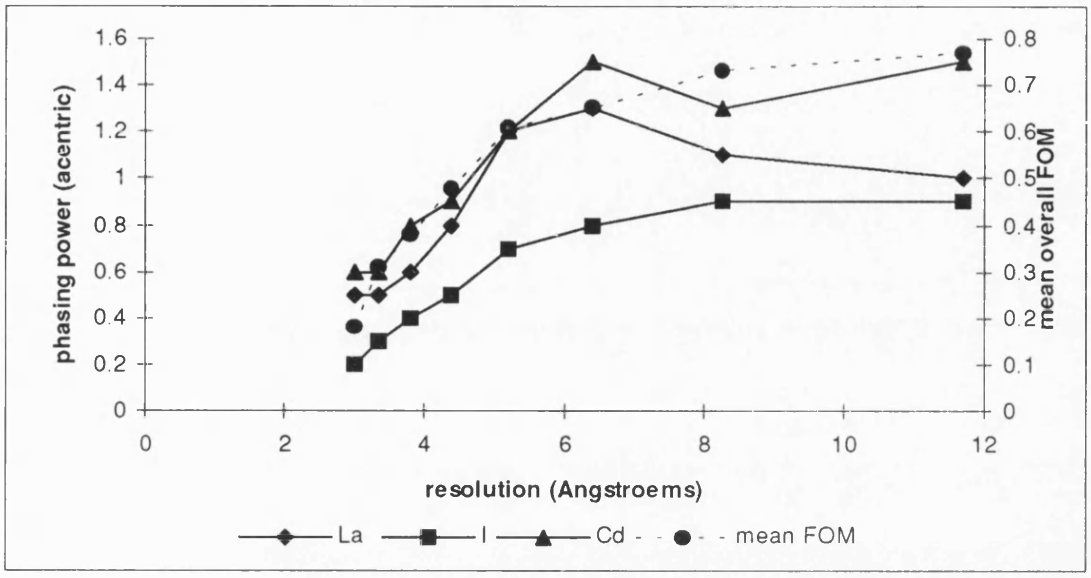
It was found that the phasing powers of the derivatives were good at low resolution, but that as the resolution limit was extended the phasing powers of the derivatives dropped off sharply. For example, at 6Å the Lanthanum derivative had a centric phasing power of 0.9 and an acentric phasing power of 1.5, whereas at 3.0Å the phasing powers were 0.2 for centric and acentric reflections. However, it was decided that since the resolution of all the derivatives extended to 3.0Å, to calculate phases in MLPHARE to 3.0Å. Table 4.4 gives the phasing statistics for all the derivatives after combination in MLPHARE. Fig. 4.5 illustrates how the phasing powers of the derivatives varied with resolution, after combination in MLPHARE. The mean overall FOM is also given as a function of resolution.

Table 4.4 Phasing statistics for the derivatives of Cadmium, Iodine and Lanthanum at 3.0Å; after combination in MLPHARE.

derivative	$CdCl_2$	KI_3	$La(ac)_3$
Heavy atom sites	C,D	B	A
Phasing power (centric, acentric) to 3.0Å	0.6, 0.8	0.3, 0.4	0.5, 0.7
Cullis R (centric, acentric) to 3.0Å	0.74, 0.42	0.93, 0.50	0.86, 0.62
Lack of closure (centric, acentric)	133.3, 106.0	109.0, 92.8	190.7, 118.8
Isomorphous difference (centric, acentric)	179.1, 126.4	117.4, 96.2	221.7, 133.5
Total no. of reflections	5354	6091	6304

Overall Figure of Merit acentric: 0.37 (5177 reflections), centric: 0.54 (1022 reflections)

Fig. 4.5 Phasing power vs. resolution for the derivatives of Iodine, Cadmium and Lanthanum, after combination in MLPHARE. The mean overall FOM is also given as a function of resolution.



Definition of terms used

$$\text{phasing power} = \text{Rms}\left\{\frac{F_H}{\text{lack of closure}}\right\} \qquad \text{Cullis R} = \frac{\text{lack of closure}}{\text{isomorphous difference}}$$

$$\text{Lack of closure} = \sum \left| \left| F_{PH} \pm F_P \right| - f_H \right| \qquad \text{Rms } F_H = \left\{ \frac{\Sigma f_H^{\,2}}{n} \right\}^{0.5}$$

$$\text{Rms Lack of closure} = \left\{ \frac{\left(\Sigma \left| \left| F_{PH} \pm F_P \right| - f_H \right|^2 \right)}{n} \right\}^{0.5}$$

$$\text{Isomorphous difference} = \sum \frac{|I_{nat} - I_{deriv}|}{\Sigma \langle I \rangle}$$

Residual = The weighted residual between observed and calculated Pattersons. The optimal value is zero.

Correlation = The weighted linear correlation between observed and calculated Pattersons. The range is 1.0 to -1.0, where 1.0 is optimal.

Mean figure of merit = mean cos(estimated phase error). The range is 0.0 to 1.0, where 1.0 is optimal.

Chapter 5

BUILDING AND REFINING THE 41kDa MODEL

Introduction

Initial phase estimates obtained from the heavy atom derivatives and structure factor amplitudes provided by the native data set now allowed the generation of an electron density map into which the model could be built. In order to generate the optimum initial electron density map, several techniques were implemented, which will be discussed below. Since the initial phase estimates will be poor due to errors such as lack of closure (Appendix 2), obtaining the best map possible is often a time consuming process. Errors introduced into the model at this stage as a result of incorrect tracing of the polypeptide backbone may be difficult to correct later on and so the model building process should be a cautious one.

The initial 3.0Å phases derived from the heavy atom derivatives were extended up to 2.5Å using solvent flattening, histogram matching and Sayre's equation in the program SQUASH (Cowtan & Main, 1993). However, the phase extended map produced by SQUASH did not have significant improvements over the original map calculated using the heavy atom phases (MIR map). In addition, phase extended maps generated using SQUASH in later stages of the refinement did not show significant

improvements over the corresponding unextended maps. The generation of a 3.0Å solvent flattened map, using the WANG algorithm (Wang, 1985), improved the density in certain core regions of the map, aiding map interpretation.

There are 4 distinct stages in the process of model building:

1. Defining the molecular envelope.
2. Generating a skeletal representation of the electron density; illustrating the path of the Cα chain.
3. Fitting the Cα positions to the electron density.
4. Insertion of the amino acid sequence into the electron density.

Interpreting the overall fold

To date, all of the neuraminidase structures known have the same overall fold. In spite of a relatively low sequence identity of 26.9% between *V.cholerae* and *S.typhimurium* (Crennell *et al.*, 1994) neuraminidase, the neuraminidase domains have a very similar overall fold. It was therefore likely that the structure of the 41kDa *M.viridifaciens* neuraminidase would strongly resemble that of the neuraminidase from *S.typhimurium,* considering also that both were in the same class of small molecular weight neuraminidases.

All map interpretation and model building was carried out using the program 'O' (Jones *et al.*, 1991) on an Evans and Sutherland ESV10/33. The structure of the *S.typhimurium* neuraminidase was manually fitted onto the MIR electron density map. A 3Å difference Fourier map was calculated from inhibitor (DANA) complex data (this was later extended to 2Å). The asymmetric shape of the inhibitor was evident and this

aided the positioning of the homologous *S.typhimurium* neuraminidase-inhibitor complex model onto the MIR map. Rigid body fitting of the structure was also performed, but did not improve the fit to the density. A skeletonized representation of the electron density map was generated using the program BONES and displayed using 'O'(O_bones)(Jones *et al.*,1991). The BONES skeleton is classified into probable main chain, possible main chain and side chain, based on length and connectivity of the fragment. The Cα backbone of *S.typhimurium* appeared to follow the connectivity suggested by the BONES extremely well and the BONES were edited, starting at the N-terminus, in order to produce a continuous chain at the level of probable main chain using the bones_break and bones_make commands in 'O'. Deviations between the path of *S.typhimurium's* Cα chain and the path predicted by the BONES possible main chain were most apparent in the loops. It was also made more difficult to determine the connectivity in some of the surface loops since the solvent flattened map was cut off and the MIR map appeared disordered in places.

Assigning the Cα positions

The next step in the model building was to assign the Cα positions to the electron density. Several conserved features in the structure of *S.typhimurium* neuraminidase aided this assignment by acting as "markers". These features consisted of:

1. FRIP motif in the salmonella enzyme. The FRIP motif at the N-terminus of the *S.typhimurium* enzyme is replaced by a YRIP motif in the 41kDa *M.viridifaciens*

neuraminidase. The arginine and the proline are completely conserved across all the neuraminidases, the arginine being a conserved active site residue.

2. Asp-boxes. Asp-boxes are known to be located at spatially equivalent positions along the polypeptide chain.

3. Conserved active site residues. These are completely conserved between the *S.typhimurium* and 41kDa *M.viridifaciens* neuraminidases.

The amino acids in the *M.viridifaciens* sequence which corresponded to these "marker" residues were inserted first and then the intervening residues added. This was performed using the bones_move command to move bones atoms to more sensible locations if required and the bones_pick_Cα command to actually convert the bones atom to the Cα of the appropriate amino acid.

Completing the model

The main chain conformation was automatically generated using lego_auto_mainchain and then the side chains were automatically inserted using lego_auto_sidechain. This process takes no account of the density, but fits portions of structure based on O's database of well refined structures to a high resolution. Hence it was then necessary to go through the whole of the structure, residue by residue, using the move commands (move_zone, move_fragment and move_atom) in order to place the Cα atoms and side chains in the best positions corresponding to the density, and using the Lego_mainchain and Lego_sidechain commands to obtain the most appropriate rotamers. Now that a complete model had been obtained, the iterative process of refinement could be initiated.

Refinement of the 41kDa model

X-PLOR version 3.1 (Brunger, 1988) was used in the refinement of the 41kDa *M.viridifaciens* neuraminidase. A summary of the stages in the refinement of the 41kDa form of *M.viridifaciens* is shown in table 5.1. The theory behind the process of refinement is discussed in Appendix 2, together with a more detailed explanation behind the X-PLOR method of refinement.

Initially in the refinement procedure, all the B-factors were set to a value of 20Å^2. Then as the refinement progressed, overall B-factor refinement was performed, with the B-factors grouped in order to reduce the parameter/observation ratio. Two groups were selected, one consisting of the polypeptide backbone atoms and one consisting of the side chain atoms. Individual B-factors were refined in the latter stages of the refinement.

Round 1 of refinement

The initial model for the 41kDa form of *M.viridifaciens* neuraminidase was refined in X-PLOR at a resolution of 2.5Å. The R-factor dropped from 59.7% to 44.1% after PREPSTAGE and decreased further to 38.4% after SLOWCOOL. This gave some credence to the model, especially since the model was at the earliest stage of refinement.

Round 2 of refinement

It was decided to approach the refinement tentatively, since the quality of the phasing statistics dramatically worsened at a resolution greater than 3Å. Structure factors for the latest model were generated using SFALL (CCP4 suite of programs) and phases from the model and the heavy atom derivatives were combined using the "COMBINE" option in SIGMAA (CCP4 suite of programs). This option combines isomorphous phases, which are input as Hendrickson-Lattman coefficients ABCD, with calculated phases for the partial structure and outputs the combined phases and Fourier coefficients 2 m F_o- D F_c, where m is the figure of merit, and D is a factor dependent on the coordinate errors (Luzzati, 1952). Where regions of the initial 3.0Å MIR and WANG maps were ambiguous or poor, these were omitted from the model and the corresponding partial map generated. As the maps gradually improved, more of the complete model was incorporated in the refinement. The R-factor for the complete model dropped from 38.1% to 36.3% in PREPSTAGE and then decreased to 31.8% after SLOWCOOL. A further 5 cycles of refinement of the complete model in X-PLOR, with concomitant manual rebuilding, resulted in a final R-factor of 27.7% after SLOWCOOL. Refinement of grouped B-factors was then implemented and the R-factor dropped to 25.1%.

Rounds 3 and 4 of refinement

A new combined map at an extended resolution of 2.5Å was generated. However the map was poorer than expected and it was postulated that at this stage of

the refinement, including the inaccurate phases derived from the heavy atom derivatives had the effect of preventing significant enhancement of the map. Therefore a $2F_o$-F_c map was generated, without combining phases from the heavy atom derivatives. Those residues which lay outside the most favoured regions of the Ramachandran plot (Ramachandran & Sasisekharan, 1968) were examined closely and after manual rebuilding with refinement in X-PLOR, the R-factor after SLOWCOOL was 26.4%. Grouped B-factor refinement reduced the R-factor to 24.6%.

Round 4 of refinement was implemented as a second cycle of round 3, examining the Ramachandran outliers remaining. In round 4 however, the resolution was extended to 1.8Å and SLOWCOOL was followed by refinement of individual B-factors. Individual B-factor refinement gave an R-factor of 27.4%.

Round 5 of refinement

In order to reduce the bias from the model, OMIT maps were calculated in X-PLOR , in an 8Å sphere around those residues which were still present in unfavourable regions of the Ramachandran plot. OMIT maps exclude phase information for a specified region of the model in the generation of the new map. Examination of the resultant OMIT map may reveal mistakes made in previous map interpretations. After subsequent reinterpretation of the positions of the Ramachandran outliers, the model refined to 26.4% after individual B-factor refinement (27.2% after SLOWCOOL).

Round 6 of refinement

Although the omit maps revealed subtle changes that could be made in the orientation and position of some of the side chains, there were no clear indications as to whether any grosser changes were necessary. In addition, 8 residues still had phi, psi angles in generously allowed regions of the Ramachandran plot and 2 residues had phi; psi angles in disallowed regions of the Ramachandran plot. Therefore simulated annealing omit maps were generated, omitting a sphere of 8Å, or a stretch of residues around these remaining Ramachandran outliers. Simulated annealing omit maps differ from conventional omit maps in that a specific region of the model is omitted from the SLOWCOOL refinement. The refined model from SLOWCOOL is then used to generated the OMIT map. The simulated annealing omit maps revealed that in several places some of the amino acids were "bunched up" resulting in residues further along the polypeptide chain being one or two residues backwards or forwards from their correct position. After manual rebuilding of these residues followed by refinement in SLOWCOOL, the R-factor decreased to 26.3% and 25.8% after individual B-factor refinement.

Round 7 of refinement

The Ramachandran plot had now improved, and there were only 4 residues in generously allowed regions and one residue in the disallowed region. In order to identify whether it was necessary to shift any further residues backwards/forwards,

conventional omit maps were generated, omitting 10% of the total number of amino acid residues at a time (about 35 amino acids) over the whole structure. However these maps did not significantly aid interpretation and so a simulated annealing omit map was generated over the stretch of residues 275-291. This region had produced a poor 3D-1D score (Luthy *et al.*, 1992) (see quality of the final model) and residues Phe 270 and Glu 277 were in generously allowed and disallowed regions of the Ramachandran plot respectively. The simulated annealing omit map revealed that following the catalytic residue Arg 276, the residues needed shifting one residue further forwards, up to Ser 289; the start of the fourth Asp-box. The resulting model refined to an R-factor of 25.0% in SLOWCOOL and had an R-factor of 24.4% after individual B-factor refinement.

Round 8 of refinement

Two simulated annealing omit maps, omitting residues 210-225 and 225-248 clearly showed the need to shift residues His 217 to Ala 231 one further back along the polypeptide chain. This explained the presence of His 217 in the generously allowed region of the Ramachandran plot. The R-factor for the resulting model decreased by 1% to 23.4% after B-factor refinement. The resulting model now had only 4 residues in the generously allowed region of the Ramachandran plot and there were no residues in the disallowed region.

As a final check to ensure that there were no further improvements that could be made by manual manipulation of the structure, simulated annealing omit maps were

generated across the whole structure, omitting 10% (35 residues) of the total number of residues at a time. The maps did not reveal any improvements that could be made.

Round 9 of refinement

Inspection of the $2F_o$-F_c map generated using the calculated structure factors from the most recent model revealed numerous isolated regions of density, of a high sigma level. Water molecules were manually placed in suitable regions of density using the water_init and water_add commands in 'O' if they met two criteria:

a) The water molecule could make at least one hydrogen bond to a protein nitrogen atom or carbonyl oxygen atom, or another water molecule's oxygen atom.

b) There were no stereochemical clashes between the inserted water molecules and protein atoms.

In total, 151 water molecules were modelled into the $2F_o$-F_c map and the resulting model including the water molecules was refined in X-PLOR. The R-factor after individual B-factor refinement was 17.9%. A new $2F_o$-F_c map was calculated and the water molecules were meticulously checked, to establish whether the above criteria were still met. Additional water molecules were added where necessary. This model refined to give an R-factor of 17.3% after individual B-factor refinement. Finally CONTACT (CCP4 suite of programs) determined if any water molecules were making stereochemical clashes with the protein atoms or neighbouring water molecules and if any water molecules had zero hydrogen bonds. After three water molecules had been rejected using the results of CONTACT, the model contained a total of 270 water molecules.

Table 5.1 Stages in refinement of the 41kDa form of *M viridifaciens* neuraminidase. All refinement was performed with a 2σ cut-off in the reflection file, except in the final cycle of refinement, where a 0σ cut-off was used.

refinement round	R-factor (%)	resolution (Å)	temperature factor refinement
1	38.4	2.5	none
2	25.1	3.0	b-group
3	24.6	2.5	b-group
4	27.4	1.82	b-individual
5	26.4	1.82	b-individual
6	25.8	1.82	b-individual
7	24.4	1.82	b-individual
8	23.4	1.82	b-individual
9	17.3	1.82	b-individual

Quality of the final model

The structure of the 41kDa form of *M.viridifaciens* neuraminidase has been determined to a resolution of 1.82Å and the crystallographic R-factor for all reflections between 20 and 1.82Å is 17.3%. The final structure consists of residues 47 to 404 and 270 water molecules have been included in the solvent model for the refined structure. Table 5.2 summarizes the refinement statistics for the final model.

Several methods were used to assess the quality of the refined model. Fig. 5.1 shows how the refined 41kDa model correlates with the 2Fo-Fc electron density map. The 3D-1D score (Luthy *et al.*, 1992) which gives a score for the biochemical sense of each amino acid side chain of the structure, was greater than zero for all regions of the structure (lowest score was 0.15 for residues 309 Pro and 310 Thr in a surface loop), indicating that all amino acid side chains were in favourable environments (fig. 5.2).

Table 5.2 Refinement statistics for the final model of the 41kDa form of *M.viridifaciens* neuraminidase.

number of non-hydrogen protein atoms	2701
number of water molecules	270
resolution range	6.0 -1.82Å
number of reflections	28,814
R-factor for all data	17.3%
rms deviation in : bond lengths bond angles	 0.016Å 2.57°
Average temperature factors: main chain side chain	 11.56Å^2 15.10Å^2

Fig. 5.1 A typical region of the refined 41kDa model superimposed on the 1.82Å 2Fo-Fc electron density map. The map is contoured at 1σ.

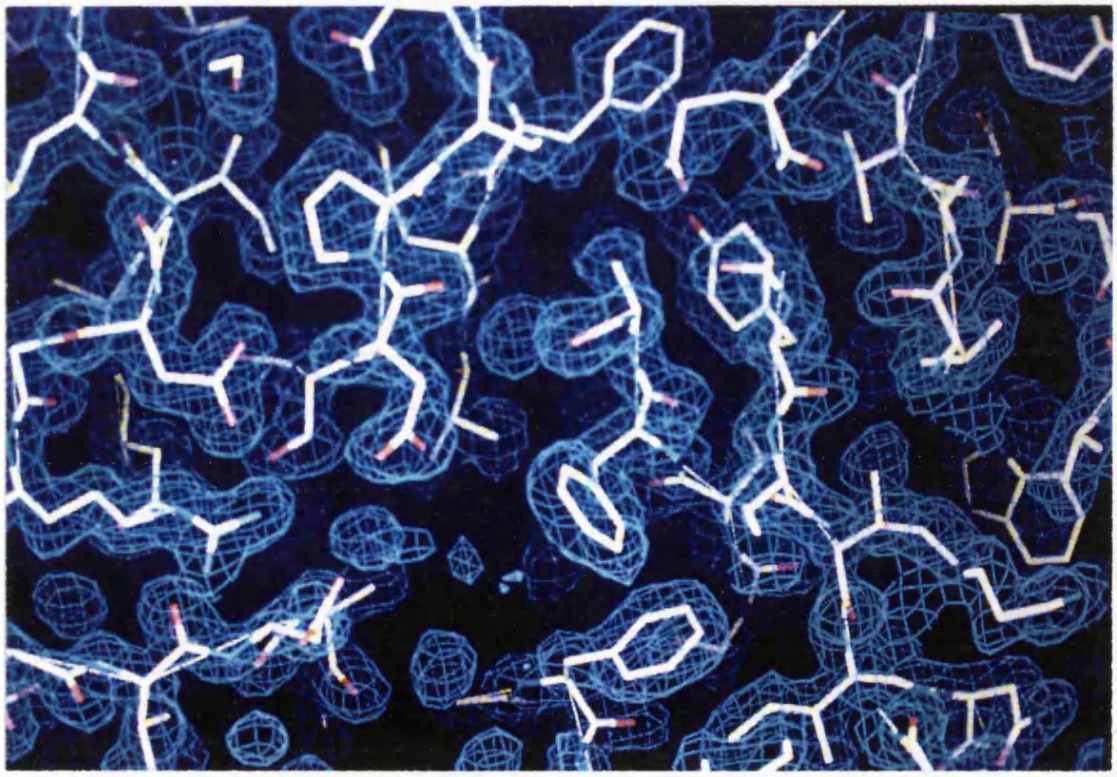
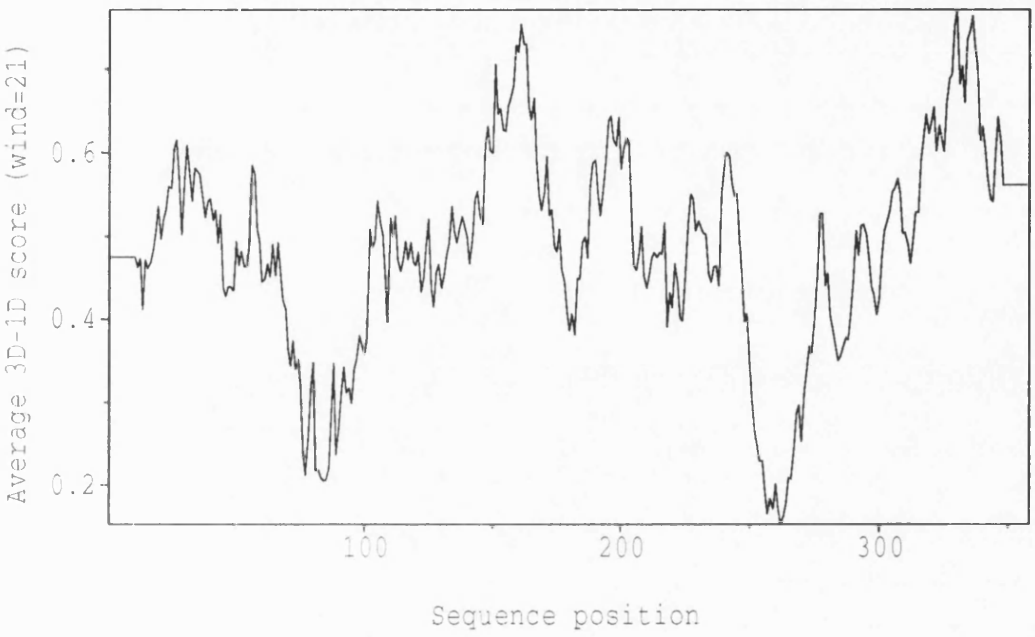
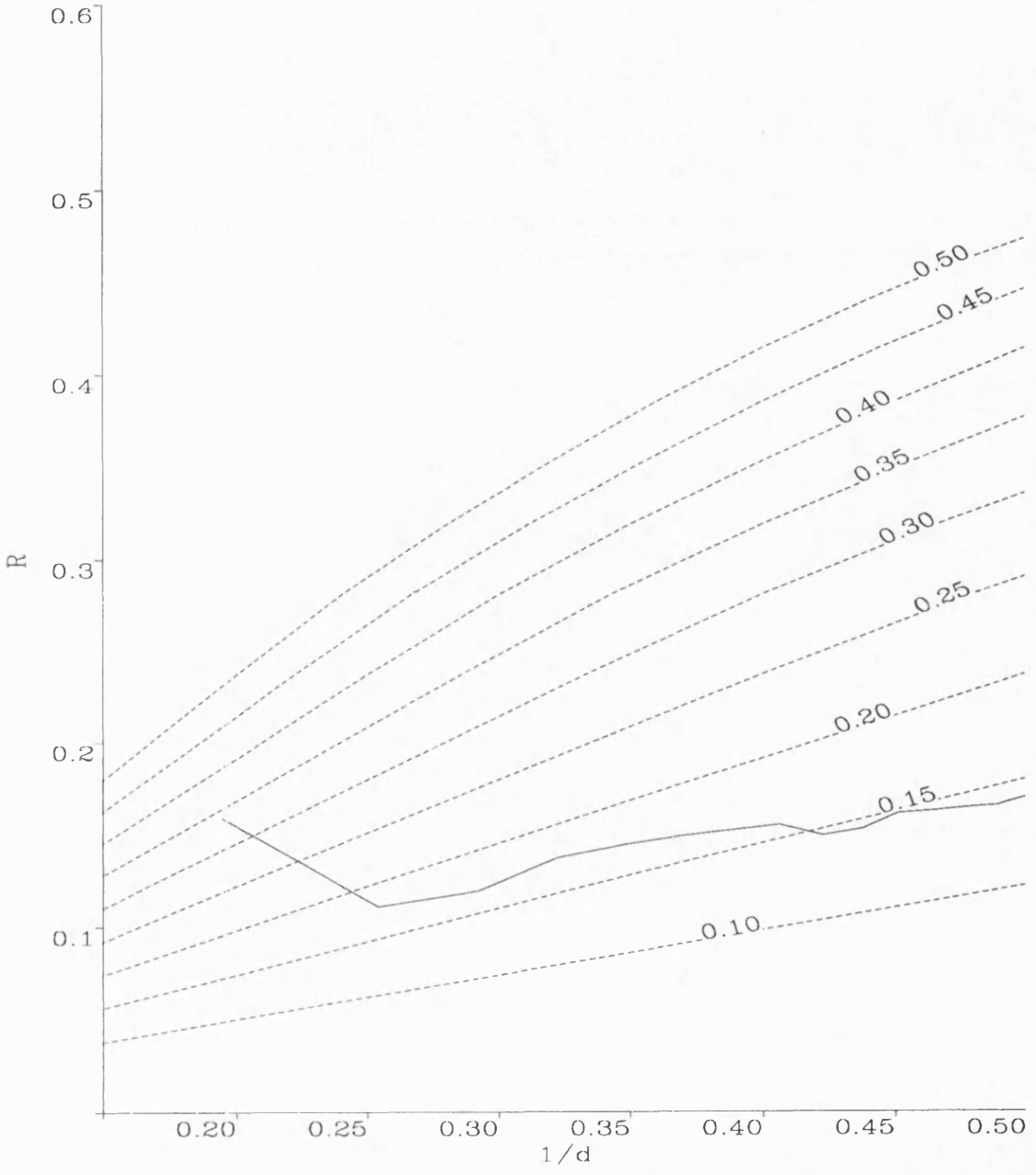


Fig. 5.2 The 3D-1D score for the final refined 41kDa model.



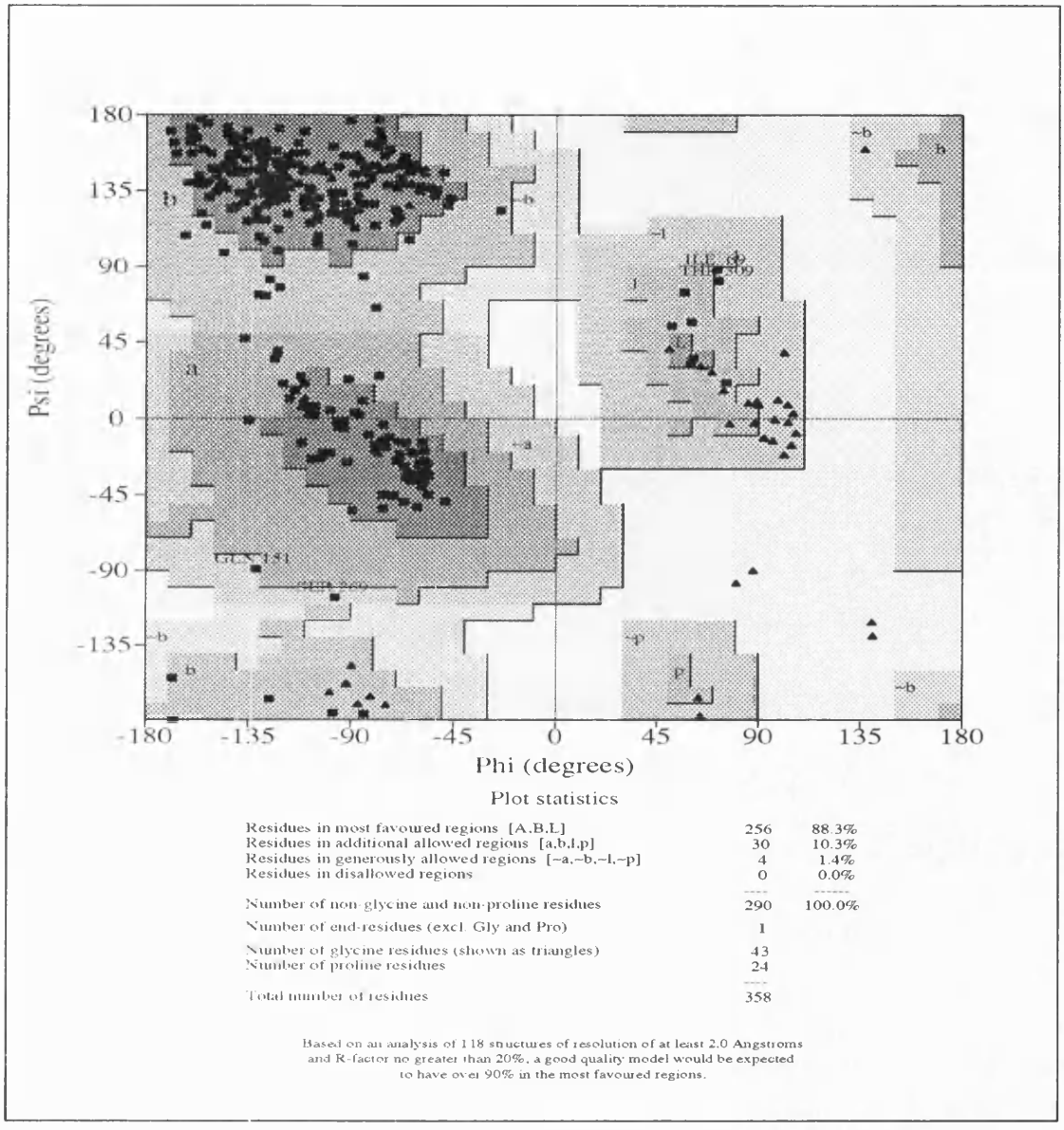
Luzatti plots (Luzatti, 1952) are useful in that they estimate mean coordinate error for a given structure. Fig. 5.3 shows a Luzatti Plot for the refined structure for data between 6 and 1.82Å with a 2σ cut-off. The plot shows R-factor versus resolution superimposed on theoretical curves that consider all discrepancies to be due to coordinate error. From this graph, the mean positional error is estimated to be between 0.15Å and 0.20Å for the refined native model.

Fig. 5.3 Luzatti plot for the refined model. The bold line gives the observed R-Factor. The broken lines represent the theoretical values associated with a given positional error in angstroems.



Procheck (Morris *et al.*, 1992) assesses the quality of the model by a variety of different methods. Procheck outputs a Ramachandran plot for the model and highlights those residues that possess ϕ, φ angles in outlying regions. As can be seen from fig. 5.4, there are no residues in the disallowed regions of the Ramachandran plot and only four in generously allowed regions, leaving 98.7% of residues in most favoured and generously allowed regions.

Fig. 5.4 The Ramachandran plot for the refined model of 41kDa *M.viridifaciens* neuraminidase.



Procheck also provides information on various main chain and side chain parameters and these are summarized in fig. 5.5. The average temperature factors for main chain and side chain atoms were 11.56Å^2 and 15.10Å^2 respectively and a plot of

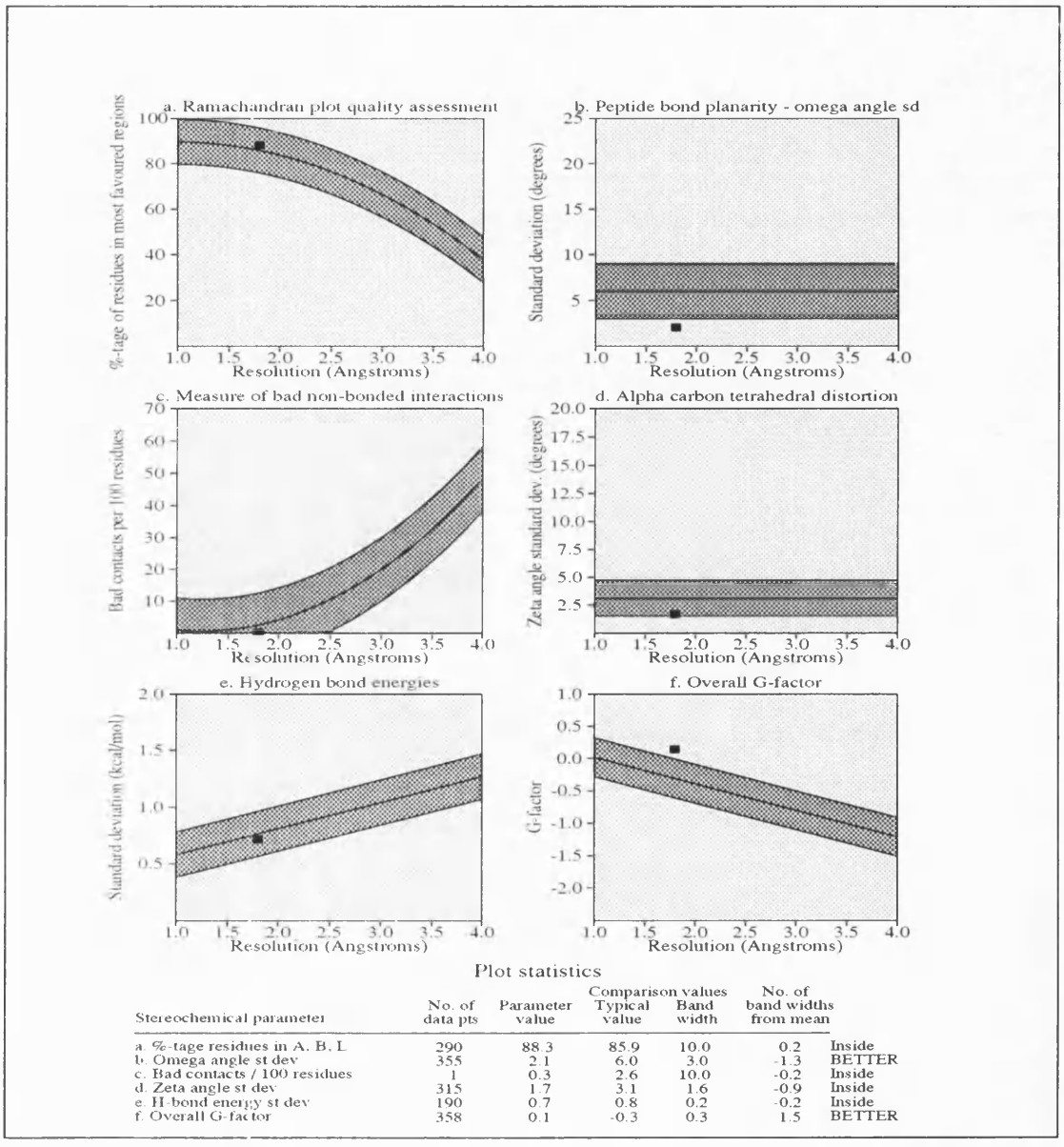
temperature factor vs. residue is given in fig. 5.6. The rms deviation in bond lengths and angles from ideality were 0.016Å and 2.57° respectively following the final cycle of refinement.

Heavy atom binding sites

The location of the heavy atom sites was investigated with respect to the refined structure (fig. 5.7). The Lanthanum heavy atom compound bound to Glu 48 and Glu 62. Glu 62 was from a separate symmetry related molecule. The Lanthanum binding site was also the binding site for the second Cadmium atom. The first binding site for the Cadmium compound was Asp 304 and Asp 302. The binding site for the Iodine atom compound was Tyr 240.

Fig. 5.5 (a) Main chain and **(b)** side chain parameters for the refined 41kDa *M.viridifaciens* neuraminidase.

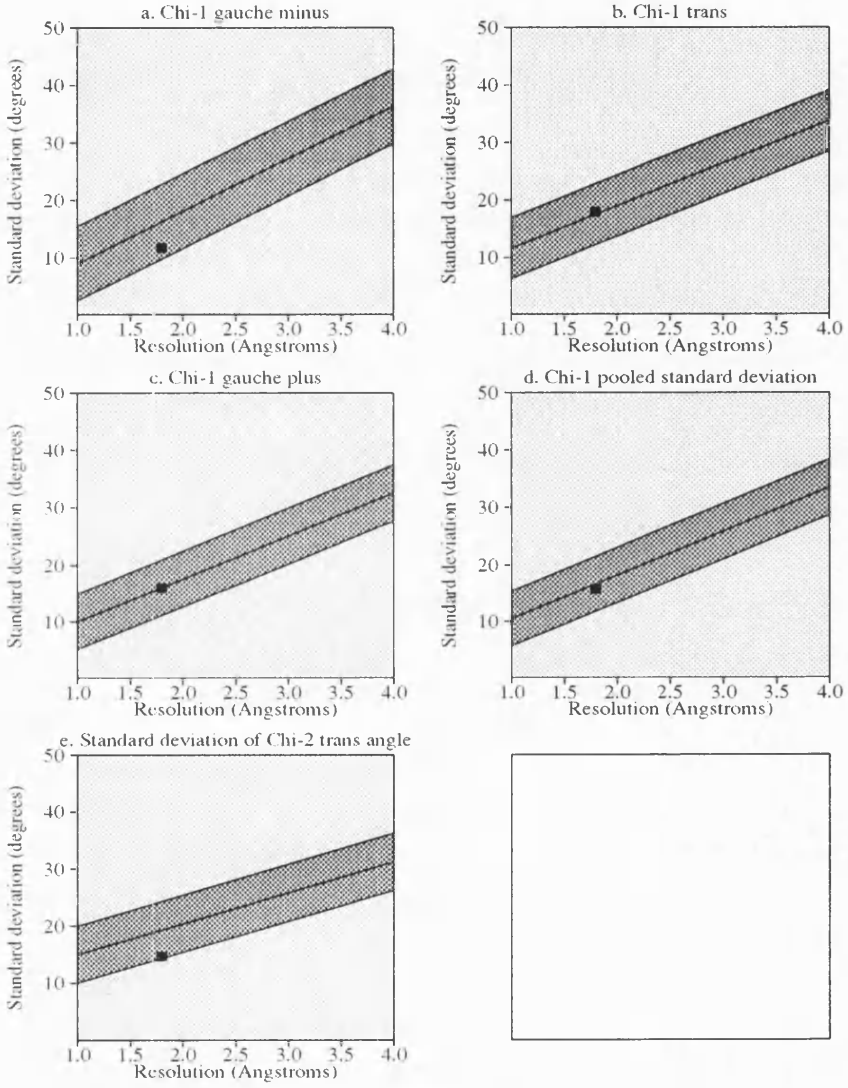
(a)



Plot statistics

Stereochemical parameter	No. of data pts	Parameter value	Comparison values Typical value	Band width	No. of band widths from mean	
a. %-tage residues in A, B, L	290	88.3	85.9	10.0	0.2	Inside
b. Omega angle st dev	355	2.1	6.0	3.0	-1.3	BETTER
c. Bad contacts / 100 residues	1	0.3	2.6	10.0	-0.2	Inside
d. Zeta angle st dev	315	1.7	3.1	1.6	-0.9	Inside
e. H-bond energy st dev	190	0.7	0.8	0.2	-0.2	Inside
f. Overall G-factor	358	0.1	-0.3	0.3	1.5	BETTER

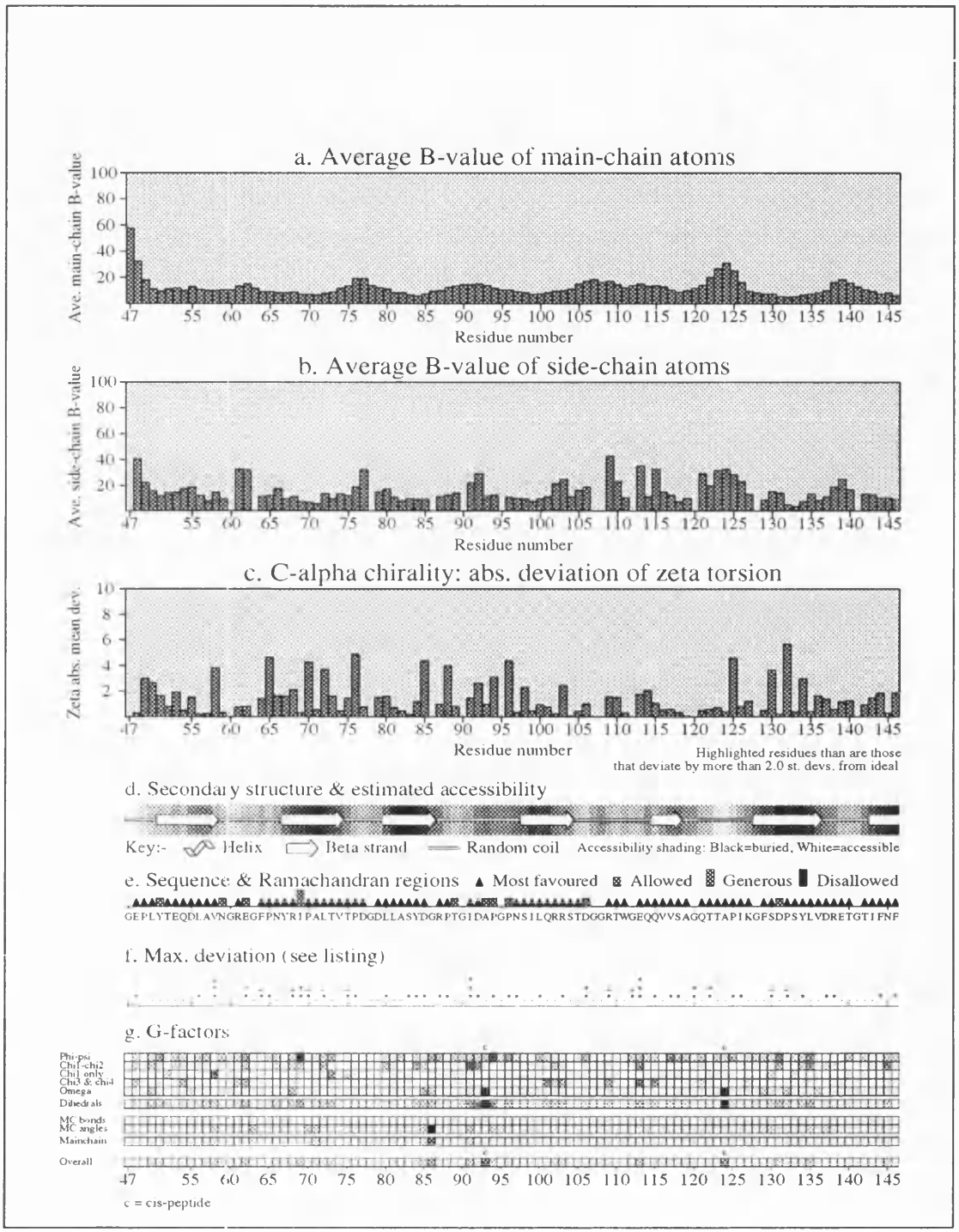
(b)

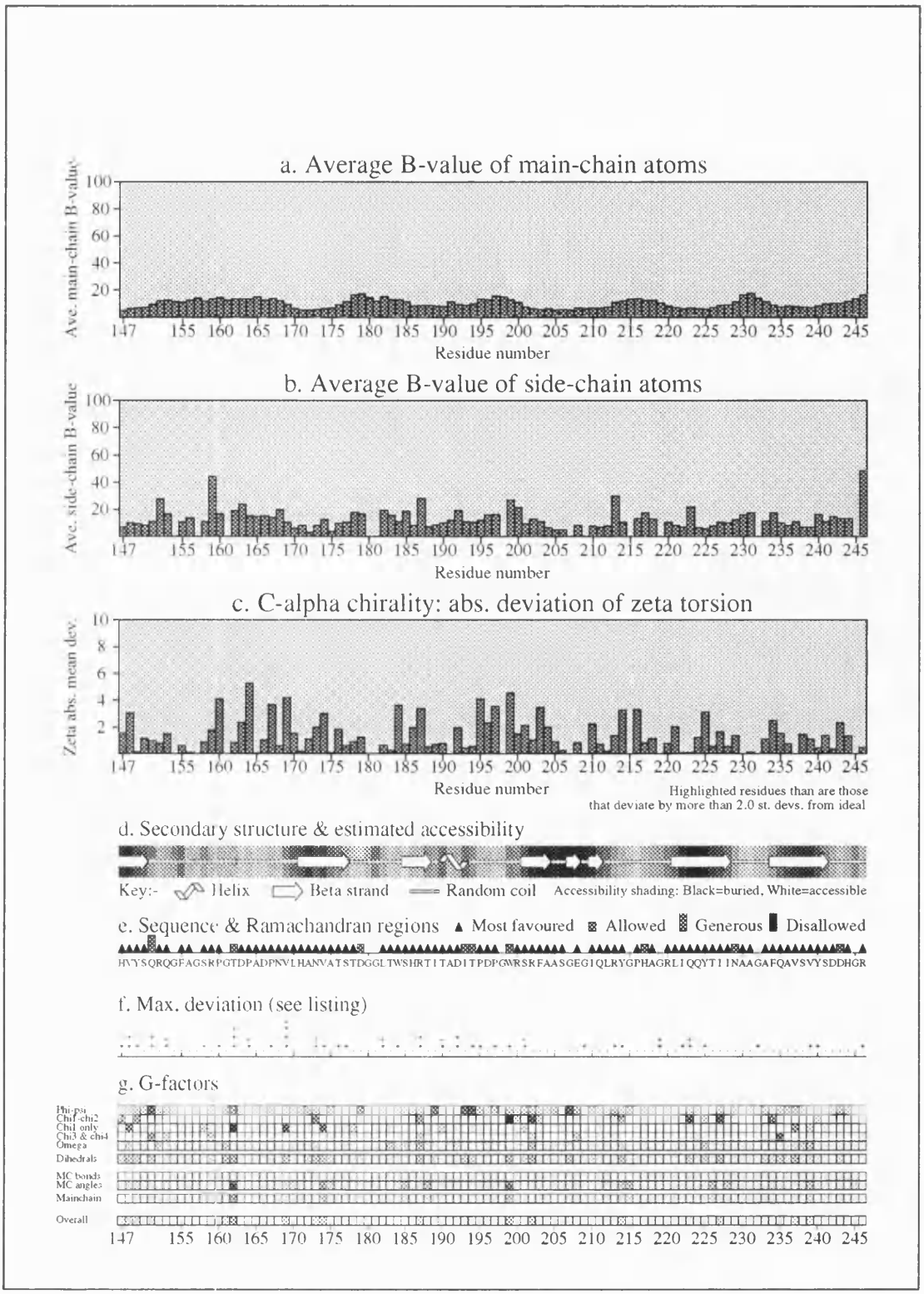


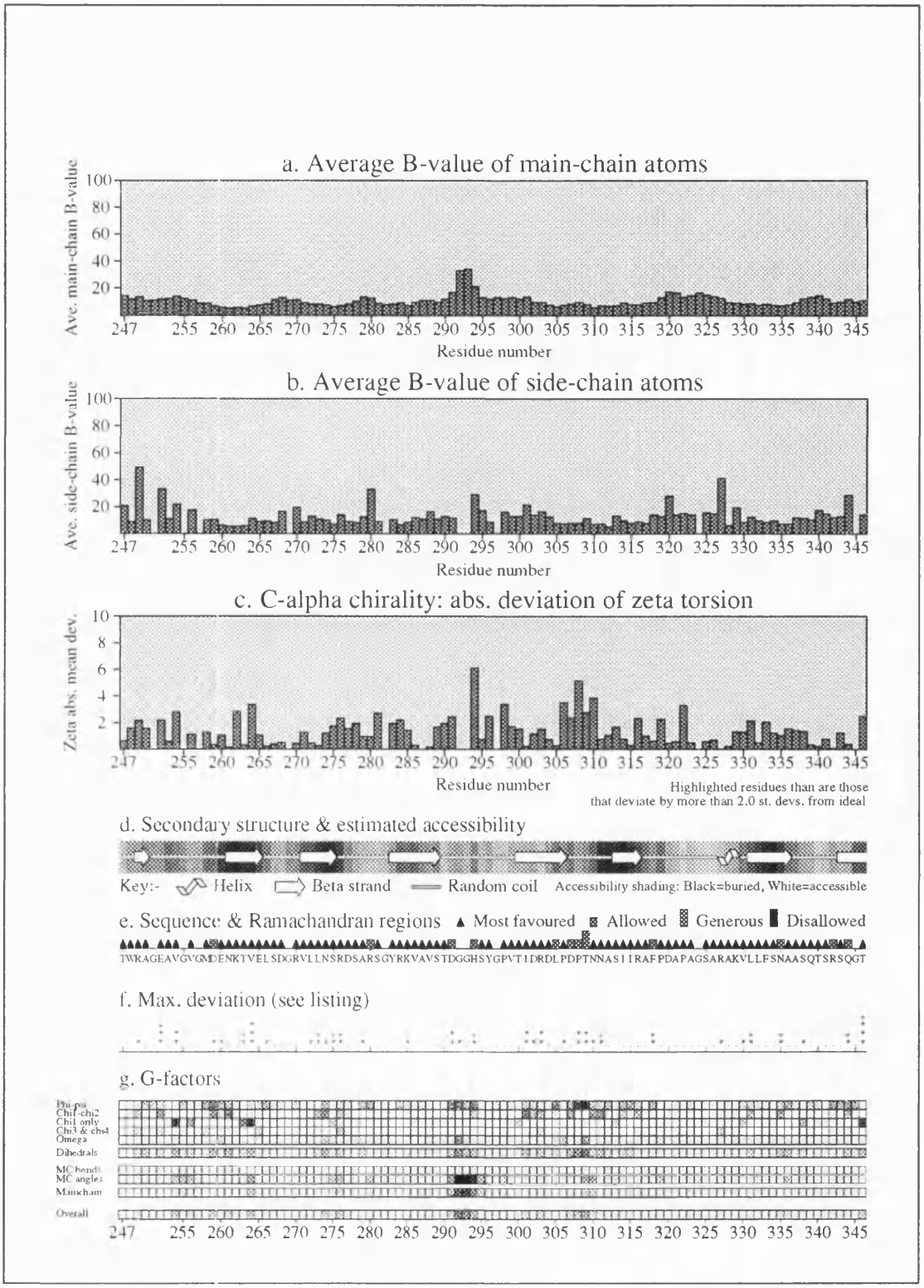
Plot statistics

Stereochemical parameter	No. of data pts	Parameter value	Comparison values Typical value	Band width	No. of band widths from mean	
a. Chi-1 gauche minus st dev	61	11.8	16.3	6.5	-0.7	Inside
b. Chi-1 trans st dev	79	18.0	17.5	5.3	0.1	Inside
c. Chi-1 gauche plus st dev	118	16.0	16.0	4.9	0.0	Inside
d. Chi-1 pooled st dev	258	15.7	16.6	4.8	-0.2	Inside
e. Chi-2 trans st dev	66	14.7	19.3	5.0	-0.9	Inside

Fig. 5.6 Temperature factor vs. residue for the refined 41kDa *M.viridifaciens* neuraminidase.




a. Average B-value of main-chain atoms
Ave. main-chain B-value
100
80
60
40
20
147 155 160 165 170 175 180 185 190 195 200 205 210 215 220 225 230 235 240 245
Residue number
b. Average B-value of side-chain atoms
Ave. side-chain B-value
100
80
60
40
20
147 155 160 165 170 175 180 185 190 195 200 205 210 215 220 225 230 235 240 245
Residue number
c. C-alpha chirality: abs. deviation of zeta torsion
Zeta abs. mean dev.
10
8
6
4
2
147 155 160 165 170 175 180 185 190 195 200 205 210 215 220 225 230 235 240 245
Residue number
Highlighted residues than are those
that deviate by more than 2.0 st. devs. from ideal
d. Secondary structure & estimated accessibility
Key:- Helix Beta strand Random coil Accessibility shading: Black=buried, White=accessible
e. Sequence & Ramachandran regions Most favoured Allowed Generous Disallowed
HVYSQRQGFAGSKPGTDPADPNVLHANVATSTDGGLTWSHRTITADITPDPGWRSKFAASGEGIQLRYGPHAGRLIQQYTIINAAGAFQAVSVYSDDHGR
f. Max. deviation (see listing)
g. G-factors
Phi-psi
Chi1-chi2
Chi1 only
Chi3 & chi4
Omega
Dihedrals
MC bonds
MC angles
Mainchain
Overall
147 155 160 165 170 175 180 185 190 195 200 205 210 215 220 225 230 235 240 245


a. Average B-value of main-chain atoms
Ave. main-chain B-value
100
80
60
40
20
247 255 260 265 270 275 280 285 290 295 300 305 310 315 320 325 330 335 340 345
Residue number
b. Average B-value of side-chain atoms
Ave. side-chain B-value
100
80
60
40
20
247 255 260 265 270 275 280 285 290 295 300 305 310 315 320 325 330 335 340 345
Residue number
c. C-alpha chirality: abs. deviation of zeta torsion
Zeta abs. mean dev.
10
8
6
4
2
247 255 260 265 270 275 280 285 290 295 300 305 310 315 320 325 330 335 340 345
Residue number
Highlighted residues than are those
that deviate by more than 2.0 st. devs. from ideal
d. Secondary structure & estimated accessibility
Key:- Helix Beta strand Random coil Accessibility shading: Black=buried, White=accessible
e. Sequence & Ramachandran regions Most favoured Allowed Generous Disallowed
TWRAGEAVGVGMDENKTVELSDGRVLLNSRDSARSGYRKVAVSTDGGHSYGPVTIDRDLPDPTNNASIIRAFPDAPAGSARAKVLLFSNAASQTSRSQGT
f. Max. deviation (see listing)
g. G-factors
Phi-psi
Chi1-chi2
Chi1 only
Chi3 & chi4
Omega
Dihedrals
MC bond
MC angles
Mainchain
Overall
247 255 260 265 270 275 280 285 290 295 300 305 310 315 320 325 330 335 340 345

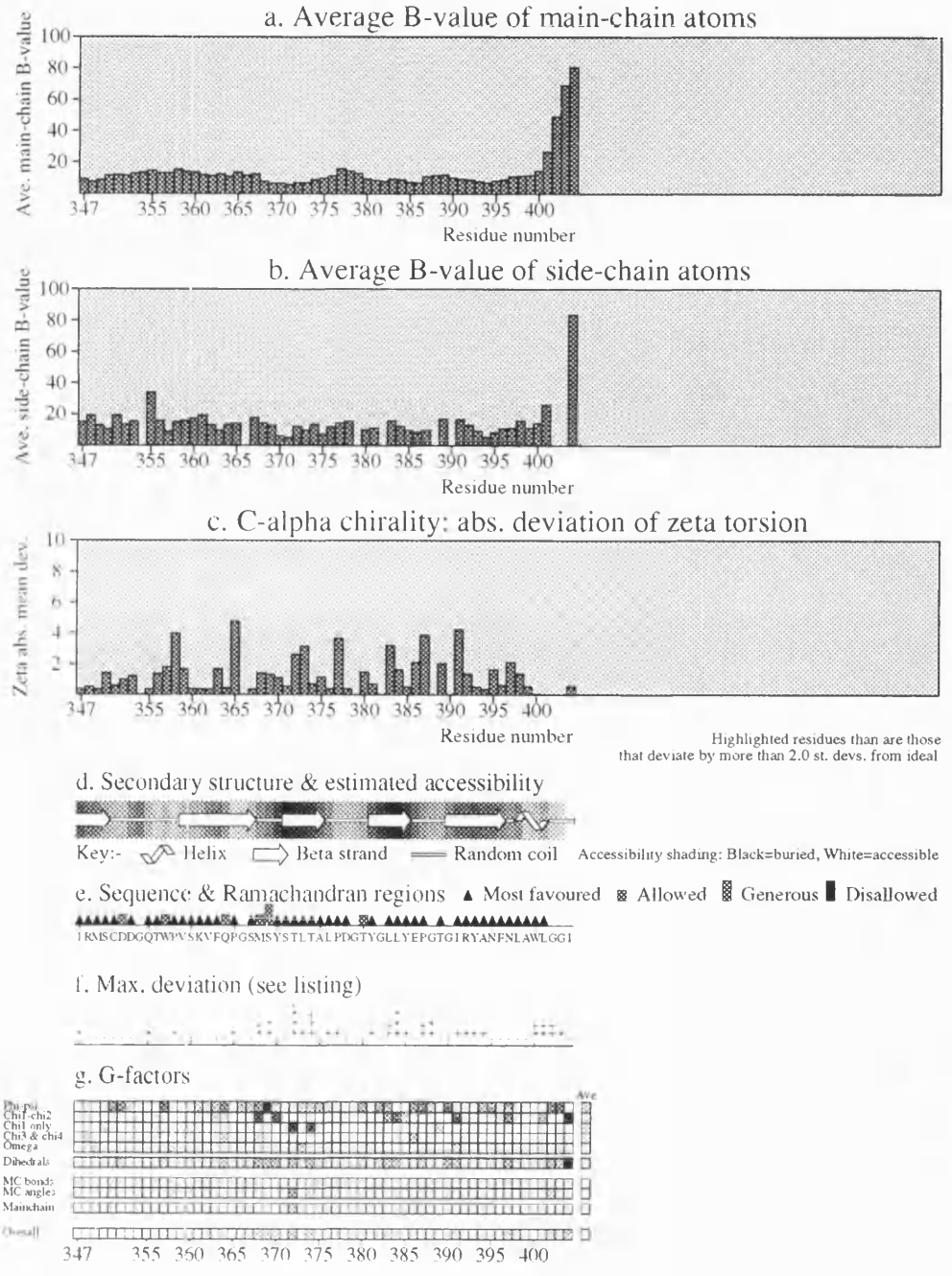
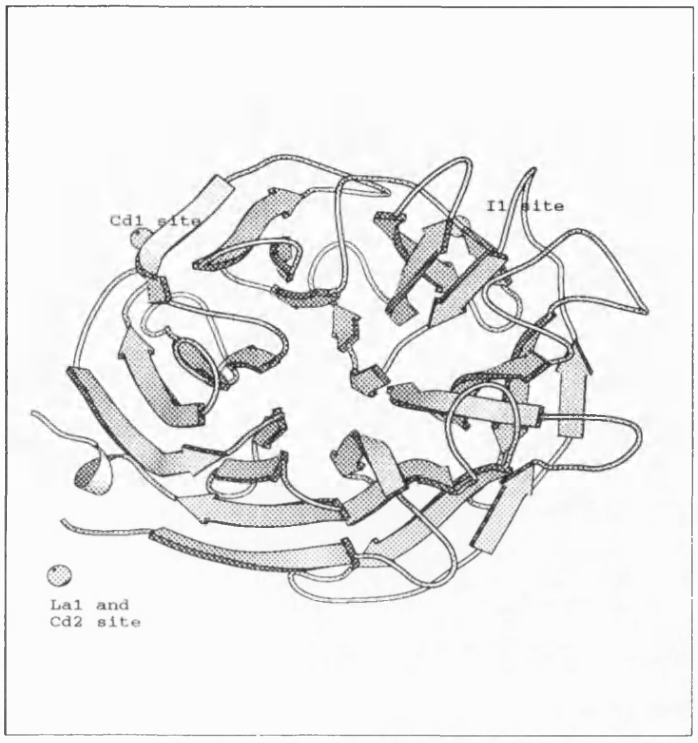

a. Average B-value of main-chain atoms
Ave. main-chain B-value
100
80
60
40
20
347 355 360 365 370 375 380 385 390 395 400
Residue number
b. Average B-value of side-chain atoms
Ave. side-chain B-value
100
80
60
40
20
347 355 360 365 370 375 380 385 390 395 400
Residue number
c. C-alpha chirality: abs. deviation of zeta torsion
Zeta abs. mean dev.
10
8
6
4
2
347 355 360 365 370 375 380 385 390 395 400
Residue number
Highlighted residues than are those
that deviate by more than 2.0 st. devs. from ideal
d. Secondary structure & estimated accessibility
Key:- Helix Beta strand Random coil Accessibility shading: Black=buried, White=accessible
e. Sequence & Ramachandran regions ▲ Most favoured ▩ Allowed ▩ Generous ▮ Disallowed
1 RAISCDDGQTWPVSKVFQPGSMSYSTLTALPDGTYGLLYEPGTGIRYANFNLAWLGGI
f. Max. deviation (see listing)
g. G-factors
Phi-psi
Chi1-chi2
Chi1 only
Chi3 & chi4
Omega
Dihedrals
MC bonds
MC angles
Mainchain
Overall
Ave
347 355 360 365 370 375 380 385 390 395 400

Fig. 5.7 Location of heavy atom sites in the 41kDa form of *M.viridifaciens* neuraminidase. Cd1 is the first Cadmium heavy atom site, Cd2 the second Cadmium heavy atom site, La1 is the Lanthanum site and I1 is the Iodine site.



Chapter 6

MOLECULAR REPLACEMENT

Introduction

The method of deriving starting phases in order to solve the electron density equation (Chapter 4) using a known structure, is molecular replacement. If the structure of a new protein is suspected to be similar to that of a known structure (e.g. from sequence similarity), then the known structure may be used as a phasing model for determining the structure of the new structure. The phases of atomic structure factors and hence of molecular structure factors, depend on the location of atoms in the unit cell. So in order to use a known protein as a phasing model, it is necessary to perform a six dimensional search in the unit cell of the protein with unknown structure, in order to determine the orientation and position of the known structure that would give phases most like those of the new protein. In order to save on computing time, the search is split into two parts; orientation (the rotation function) and the position of the search model (translation function). The theory behind the use of the rotation and translation functions in molecular replacement will be discussed in Appendix 2.

The Strategy

Since the structure of the 41kDa form of *M.viridifaciens* neuraminidase was known, a potentially high quality search model existed for solution of the

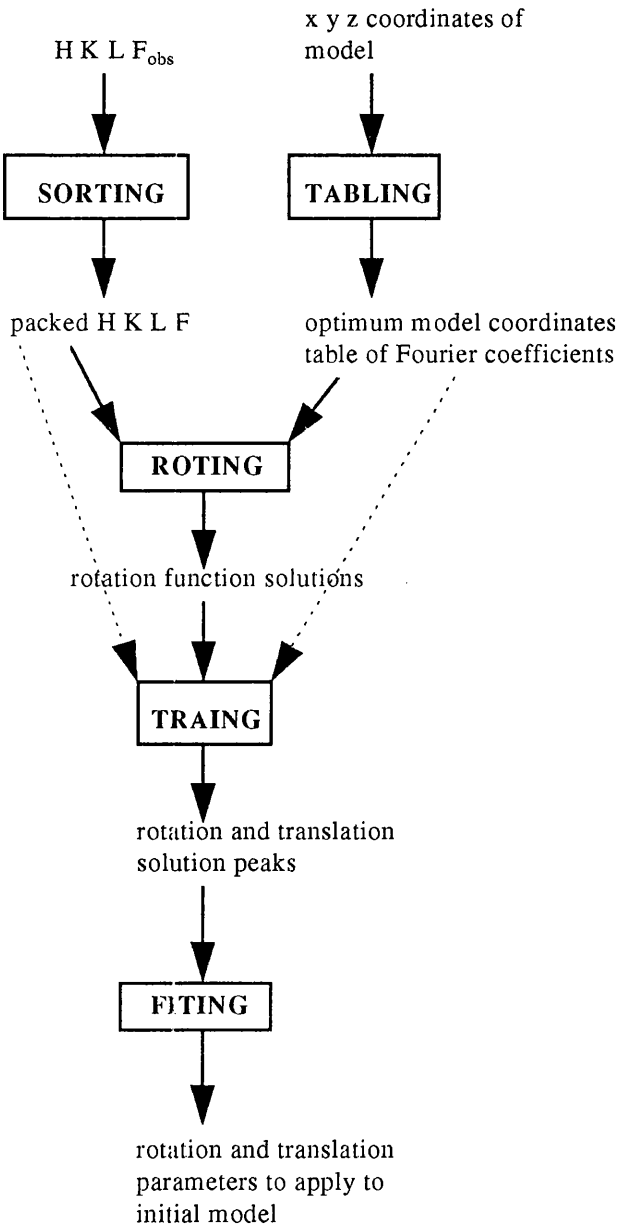
68kDa form by molecular replacement. Initially, the only data set available for the 68kDa crystals was 93.5% complete (>1σ) to 6Å and hence the rotation and translation functions were performed with a high resolution limit of 6Å. The molecular replacement package AMoRE (Navaza, 1992) was used for all calculations.

AMoRE

The molecular replacement package AMoRE generates structure factors for the model and enables the search of a large number of rotation function peaks, as well as the calculation of the corresponding translation functions. AMoRE also applies rigid body refinement to the solutions. The search for a molecular replacement solution in AMoRE consists of a distinct number of stages, which are summarised in fig. 6.1. The first step, SORTING, consists of packing and sorting the F_{obs} for all reflections of the type hkl obtained from the data set for the target (unknown) structure. SORTING outputs a packed file of hkl F_{obs} and $SIGF_{obs}$ in the P1 asymmetric unit. In the second step, TABLING places the search model in a small box and rotates the model coordinates such that the principal axes of inertia of the model are parallel to the box axes and translates the coordinates so that the centre of mass is at the origin. This ensures an optimum placement of the model in order to begin a rotation search. TABLING outputs the new model coordinates and prepares a table of continuous Fourier coefficients from the coordinate model. The rotation function is carried out in ROTING (Crowther, 1972; Navaza, 1987). Firstly, ROTING calculates structure factors for the model in a suitable cell, using the Fourier coefficients obtained from TABLING and outputs a set of packed

structure factors, in the same format as the output from SORTING. ROTING then calculates spherical harmonics for both the crystal and the model, which are required for calculation of the rotation function. Any promising rotation function output solutions can then be put into the TRAING stage of AMoRE. TRAING calculates the translation function for a given rotation function output solution, using the Crowther and Blow overlap function method (Crowther & Blow, 1967). Fast rigid body refinement of any likely rotation and translation peaks can then be carried out in FITING, (Castellano *et al.*, 1992). If molecular replacement has been successful, then the refined rotation and translation matrix produced by FITING, which takes into account the rotation and translation of the search model in TABLING, can be applied to the original search model.

Fig. 6.1 Flow chart summarising the stages involved in molecular replacement using AMoRE.



A Molecular replacement solution for the 68kDa form of *M.viridifaciens* neuraminidase using AMoRE

A single set of parameters was used in AMoRE, since a molecular replacement solution was obtained immediately. The resolution limits of the data used for the target structure were between 100 and 6Å. Structure factors were generated to 4Å in TABLING, the Shannon rate for sampling the coordinate map was 2 and the scalefactor which determines how finely the model structure samples reciprocal space was 4. The sphere radius used in ROTING, within which interesting intermolecular vectors should lie, was between R_{min} = 0 and R_{max} (Patterson cut off radii) = 20Å. Navaza suggests that this radius of integration should not be greater than the maximal distance from the centre of mass, 31.46Å as calculated by TABLING. The resolution range over which the cross rotation function was performed, was between 12 and 6Å. A list of the top ten rotation function peaks are listed in Table 6.1.

Table 6.1 The top ten rotation function peaks obtained in AMoRE. Symmetry related peaks are omitted for clarity (1σ = 8.1).

α	β	γ	Correlation Coefficient
261.28	100.13	159.18	32.4
98.50	25.00	33.21	29.8
299.74	51.11	10.10	29.5
128.69	87.09	189.36	28.7
150.39	53.74	322.67	28.3
107.04	115.56	53.93	28.1
185.60	87.84	196.98	27.9
104.09	95.16	268.00	27.4
8.84	40.91	137.04	26.5
98.32	103.83	227.65	25.4

The peak with the highest correlation coefficient and lowest R-factor is generally the correct solution. This peak had an AMoRE correlation coefficient of 32.4 (4.0σ), the next highest peak having a correlation coefficient of 29.8 (3.7σ).

The rotation function peaks were edited to remove the symmetry related peaks and put into TRAING, in order to find the corresponding translation function peaks. The translation function search was again performed with resolution limits between 12 and 6Å, and a Shannon rate of 2. The top ten best translation function peaks obtained, corresponding to the above rotation function peaks are listed in table 6.2 below. Note that since translations along the b axis are arbitrary in space group $P2_1$, all b translations are zero.

Table 6.2 The top ten rotation and corresponding best translation function peaks obtained in AMoRE. Symmetry related peaks are again omitted for clarity.

α	β	γ	a	b	c	correlation coefficient	R-factor
261.28	100.13	159.18	0.11111	0.00000	0.38095	30.1	47.7
98.50	25.00	33.21	0.05556	0.00000	0.04762	19.3	51.7
299.74	51.11	10.10	0.05556	0.00000	0.07143	16.0	53.1
128.69	87.09	189.36	0.20833	0.00000	0.34524	20.3	51.0
150.39	53.74	322.67	0.36111	0.00000	0.09524	18.0	52.3
107.04	115.56	53.93	0.04167	0.00000	0.19048	14.5	51.7
185.60	87.84	196.98	0.30556	0.00000	0.26190	20.9	50.4
104.09	95.16	268.00	0.01389	0.00000	0.32143	20.3	51.2
8.84	40.91	137.04	0.26389	0.00000	0.13095	18.4	53.1
98.32	103.83	227.65	0.01389	0.00000	0.11905	24.6	49.6

The highest peak resulting from the translation function had an AMoRE correlation coefficient of 30.1 (3.3σ). The next highest peak (not shown - for the same rotation function solution) had a correlation coefficient of 25.1 (2.9σ).

Having found a likely candidate for a molecular replacement solution, the top peak from the translation function was rigid body refined and this gave a correlation coefficient of 48.5 and a crystallographic R-value of 41.5%. The final rigid body refined orientation and position of the model is given in table 6.3:

Table 6.3 Rigid body refined orientation and position of the 41kDa *M.viridifaciens* neuraminidase model, which has been re-orientated by AMoRE.

α	β	γ	a	b	c	correlation coefficient	R-factor
254.02	104.89	156.08	0.10	0.00	0.38	48.5	41.5

The search model was rotated in TABLING so that its principle axes of inertia were parallel to the box axes and translated so that the centre of mass is at the origin. Hence, the final stage in FITING is to convert the solution in table 6.3, so that it applies to the original model. The rotation and translation that was finally applied, is shown in table 6.4.

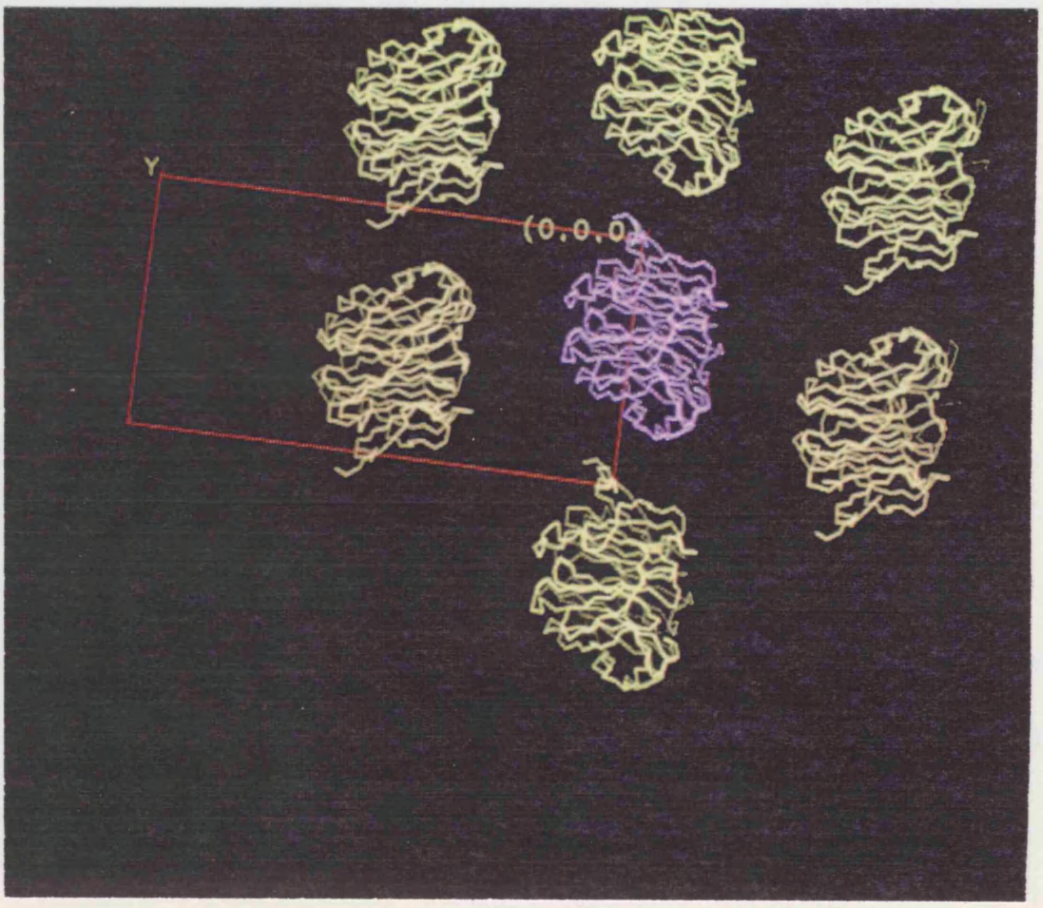
Table 6.4 Orientation and position of the original coordinates of the 41kDa portion in the 68kDa *M.viridifaciens* neuraminidase asymmetric unit.

α	β	γ	a	b	c	correlation coefficient	R-factor
32.97	54.29	217.94	-56.42	0.00	14.56	48.5	41.5

The molecular replacement solution for the position of the 68kDa form of *M.viridifaciens* neuraminidase was further verified by manual examination of packing between symmetry related molecules in the unit cell (Fig. 6.2), using the graphics package "O". No clashes, or bad contacts between symmetry related

molecules were apparent and this solution was subsequently used to solve the structure of the 68kDa *M.viridifaciens* neuraminidase.

Fig. 6.2 Molecular packing of the 41kDa *M.viridifaciens* neuraminidase search model in the 68kDa *M.viridifaciens* neuraminidase unit cell.



Discussion

The ease with which a molecular replacement solution was found for the 68kDa form of *M.viridifaciens* neuraminidase reflects two major advantages in the way in which the search for a solution was conducted. Firstly, the search model

was known with almost complete certainty to represent 60% of the target structure and therefore represented a high quality search model. Secondly, the molecular replacement package AMoRE is recognized as a powerful tool in molecular replacement and has been employed successfully in a number of cases where standard packages have failed.

Chapter 7

BUILDING AND REFINING THE 68kDa MODEL

Rigid body refinement

The first step after determining the rigid body refined molecular replacement solution to the 41kDa *M.viridifaciens* neuraminidase search model was to perform a rigid body refinement of the rotated and translated model, implemented in X-PLOR (round 1 of refinement). All subsequent refinement was also performed using X-PLOR version 3.1. The refinement procedure is summarized in table 7.1. The free R-factor (Brunger, 1992) decreased from an initial value of 46.3% to 44.5% at a resolution of 3.5Å. Rigid body refinement uses the complete energy function E_{TOTAL}, (Appendix 2), minimizing the six rotational and translational degrees of freedom for specified groups of atoms.

Generating the first electron density map

Using phases derived from the rigid body refined 41kDa *M.viridifaciens* neuraminidase model, a difference Fourier map of the type $2F_o$-F_c was generated. Firstly, structure factors for the model were generated using SFALL. A scale factor was then applied to the observed and calculated structure factors in

RSTATS (CCP4 suite of programs), in order to put both on the correct scale. Using the calculated phases from the partial structure information of the model, the "Partial" option in SIGMAA was used to write out weighted Fourier coefficients for a difference map of the type $2F_o\text{-}F_c e^{i(\alpha calc)}$. The fast Fourier program FFT (CCP4 suite of programs) was then used to calculate the difference Fourier with the above coefficients and generate the electron density map at a resolution of 3.5Å. The map was also solvent flattened using the WANG procedure and this greatly improved the quality of the map.

Inspection of the 1st electron density map

The fit of the 41kDa *M.viridifaciens* neuraminidase model to the 1st map was excellent and the correct side chain density was apparent (fig. 7.1). In addition, extra density due to the remaining 27kDa of the structure was revealed (fig. 7.2).

Fig. 7.1 A typical region of the initial solvent flattened difference electron density map with the 41kDa *M.viridifaciens* neuraminidase model superimposed on the density. The resolution of the map is 3.5Å and is contoured at 1σ.

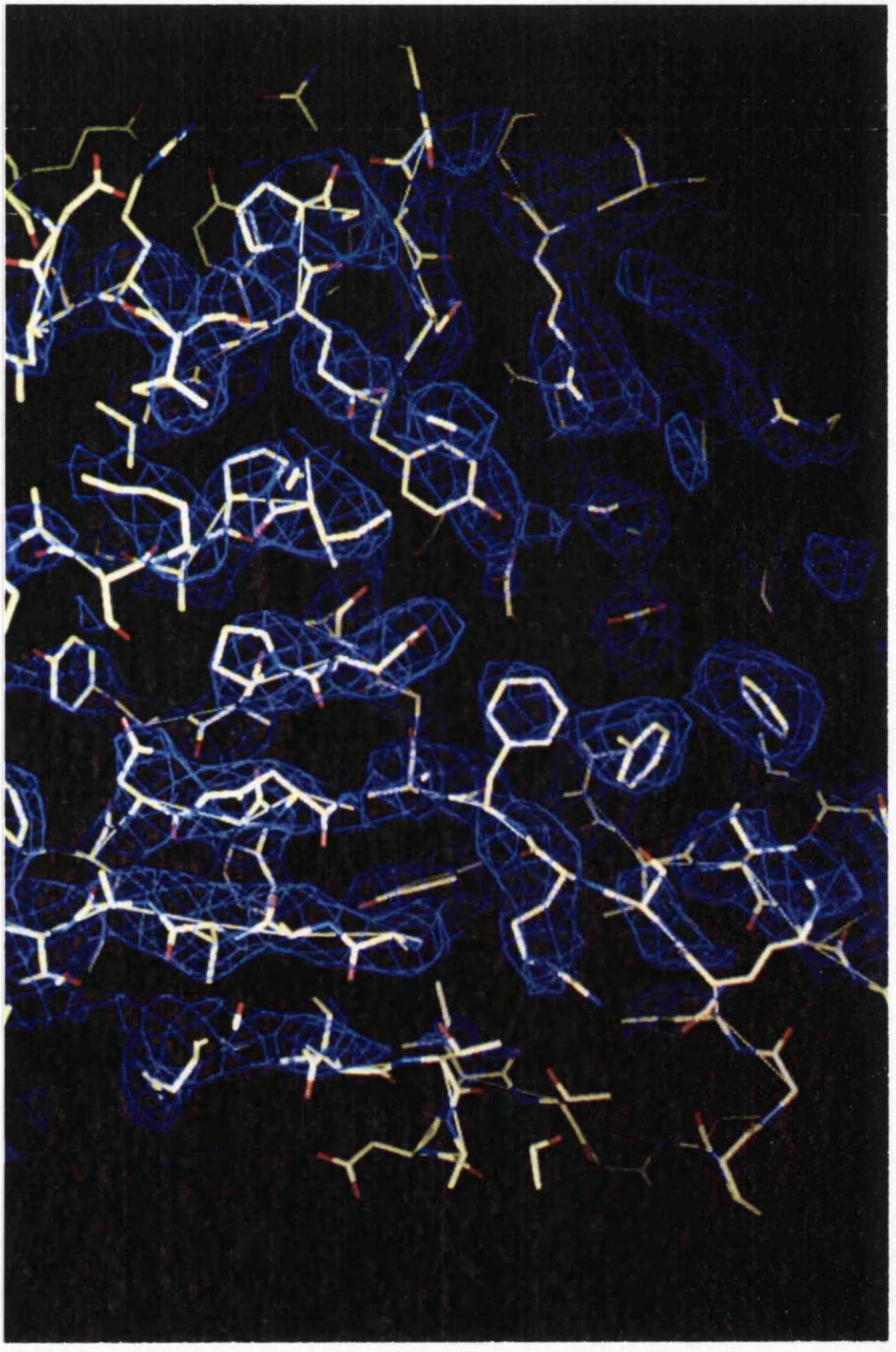
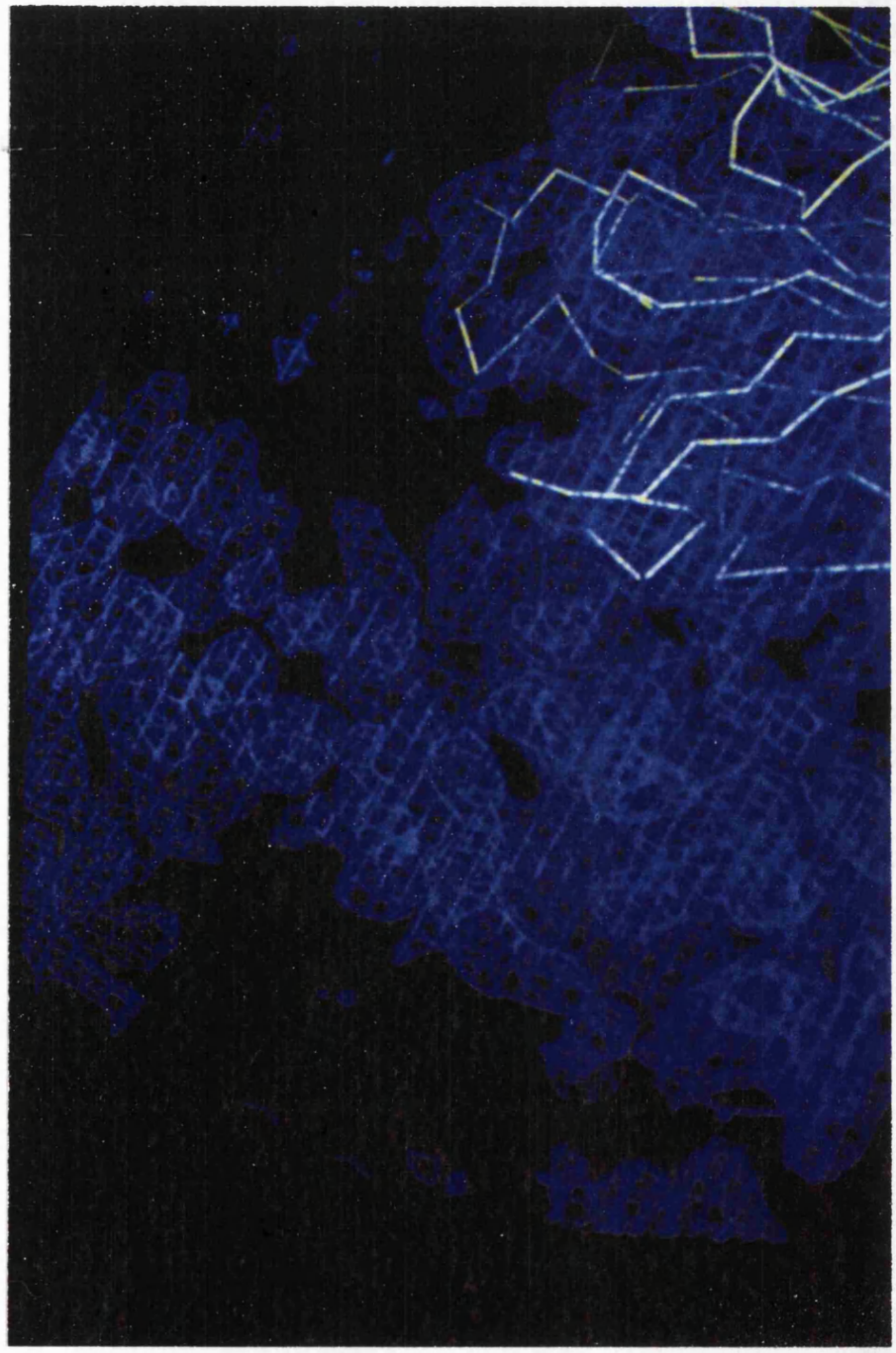


Fig. 7.2 A view of the same map, showing the extra density corresponding to the additional 27kDa of the 68kDa structure.



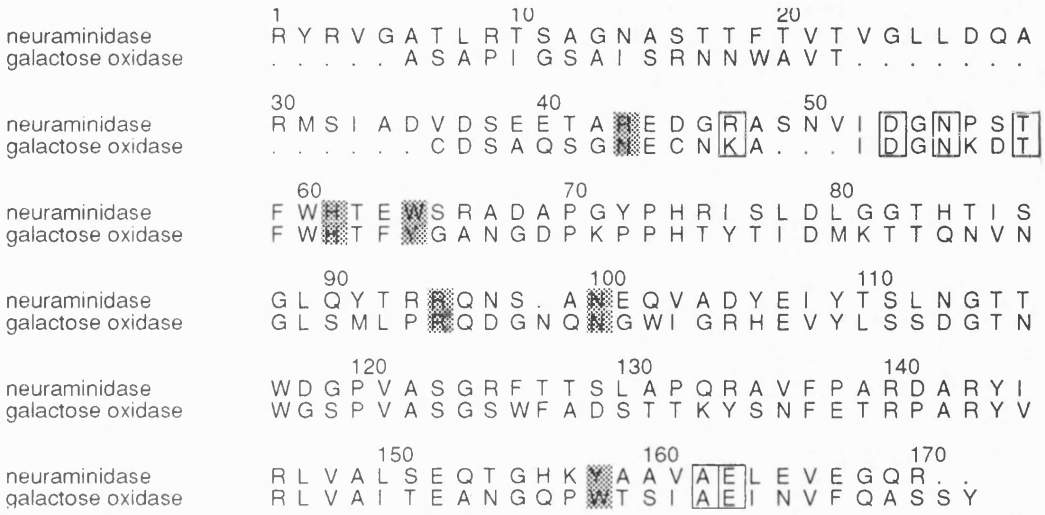
The extra density at the C-terminus of the 41kDa domain revealed the path that the α-carbon chain would take into the next domain and the density appeared to indicate the presence of two domains. However the data set for the 68kDa *M.viridifaciens* neuraminidase was only 57% complete to 3.5Å and this was reflected in the overall quality of the map. The density could only be clearly interpretated for several residues after the C-terminal Ile of the 41kDa protein and an overall fold for the extra 27kDa was not apparent. In order to trace the polypeptide backbone through the extra density, significant improvements in the map needed to be made.

Modelling the electron density for the extra 27kDa of protein

In order to improve the difference Fourier map obtained using starting phases from the molecular replacement solution, a model was required for at least part of the unknown structure, which would improve the overall phase information. To reduce the bias introduced by the 41kDa sequence, this was removed from the 68kDa sequence and the remaining 27kDa sequence searched against the SWISSPROT amino acid sequence database using FASTA in the GCG suite of software. The database search revealed a 34.6% amino acid sequence identity over a stretch of 127 amino acids with galactose oxidase from *Dactylium dendroides*, whose structure has been elucidated (Ito *et al.*, 1991, 1994). The second highest identity of 30.9% within a 139 amino acid region, lay

with the neuraminidase precursor from *C. septicum*. The sequence alignment of the 27kDa sequence against galactose oxidase is shown below in Fig. 7.3.

Fig. 7.3 Sequence alignment of the C-terminal 27kDa sequence of 68kDa *M.viridifaciens* neuraminidase (top sequence) against galactose oxidase from *Dactylium dendroides* (lower sequence). Only the region of high identity is shown. Residues involved in cation binding and galactose binding are boxed and shaded respectively (Ito *et al.*, 1991).



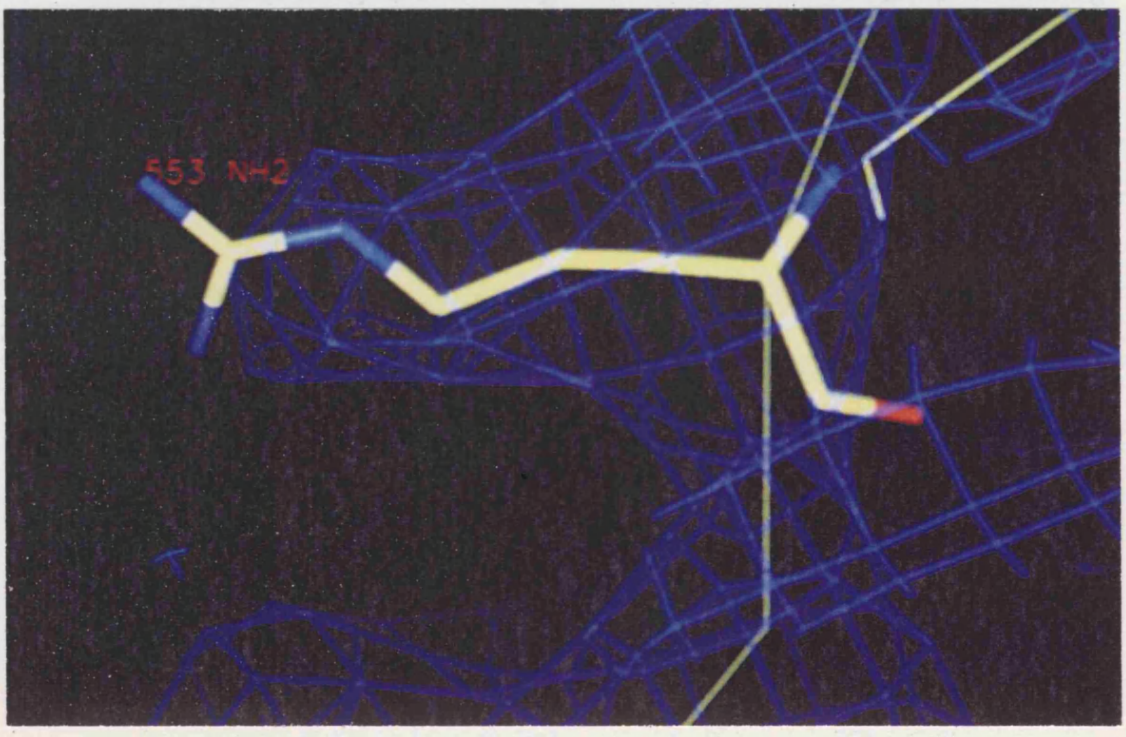
Galactose oxidase is a three domain structure, consisting almost entirely of β-strands (Chapter 9) (fig. 7.4). The first domain is a β-sandwich, consisting of a five stranded antiparallel β-sheet facing a three stranded antiparallel β-sheet. The second domain is the largest of the three and has the β-propeller fold with pseudo 7-fold symmetry. The C-terminal third domain is formed from seven β-strands, which make a complicated hydrogen bond network.

From examination of the sequence alignment, the identity with the C-terminal part of the 27kDa sequence was shown to lie within the N-terminal

domain 1 of galactose oxidase. Hence, domain 1 was moved manually onto the *M.viridifaciens* difference map using the graphics package O and was found to fit remarkably well. The galactose binding domain was then mutated to the *M.viridifaciens* sequence using the O command mutate_replace. Side chain density for the *M.viridifaciens* sequence at residues where the two sequences differed was evident, giving further credence to the use of domain 1 of galactose oxidase as a model (fig. 7.5). At this stage however, the density for the intervening domain linking the galactose binding and 41kDa domain could not be interpreted. Therefore in an attempt to improve the quality of the map, it was decided to refine the 41kDa and galactose binding domain and calculate a new map. Throughout the 3.5Å refinement the 41kDa domain was kept fixed, together with its original B factors.

Fig. 7.4 Orthogonal views of galactose oxidase from *Dactylium dendroides*. The catalytic domain is shown in mustard, the immunoglobulin domain in burgundy and the galactose binding domain in olive. A copper ion involved in catalysis by galactose oxidase is shown as a copper sphere and the sodium ion as a grey sphere.

Fig. 7.5 Region of the 3.5Å $2F_o$-F_c electron density map with the mutated domain 1 of galactose oxidase superimposed on the density. Residue Arg 553 (renumbered) in the *M.viridifaciens* sequence which corresponds to a Threonine in galactose oxidase has the correct side chain density.



Refinement and completion of building of the 68kDa *M.viridifaciens* neuraminidase

Round 2 of refinement

The galactose binding domain mutated to the *M.viridifaciens* sequence and 41kDa domain were refined in X-PLOR as rigid bodies. The free R-factor was 47.1% before and after the refinement.

Round 3 of refinement

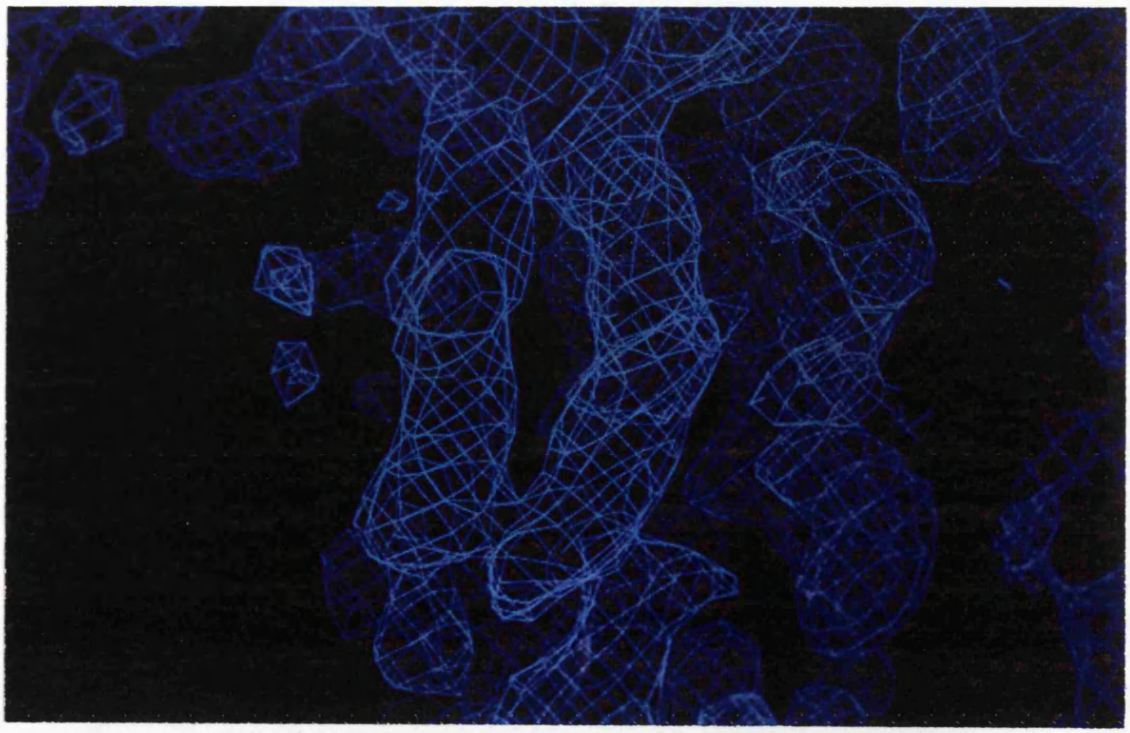
The galactose binding domain was refined using PREPSTAGE in X-PLOR at a resolution of 3.5Å. After PREPSTAGE the free R-factor for the model was 44.1%.The refined coordinates were then used to generate a new

partial map, using the same protocol as described above. The corresponding solvent flattened map was also calculated using the WANG procedure.

Round 4 of refinement

The improved phases resulting from the inclusion of the galactose-binding domain in the model, improved the quality of the map significantly and it now became apparent that the linker domain consisted of seven β-strands (Fig. 7.6). Various β-strand structures of the right dimensions were superimposed onto the electron density corresponding to the linker domain in an attempt to find a model. These included various immunoglobulins and fibronectin modules. However it was not possible to find a model that unambiguously fitted the density. Therefore the density was traced without the aid of a model, up to the end of the first β-strand. After this the density became ambiguous and so a new partial map was generated with the first strand incorporated as a polyalanine chain.

Fig. 7.6 Improved electron density map of the linker domain.



Round 5 of refinement

This process was repeated, incorporating more of the polyalanine structure as each new partial map improved. After the whole of the linker domain had been built as polyalanine, the model was refined in PREPSTAGE and a new SIGMAA weighted map calculated. In all rounds of refinement in X-PLOR, the 41kDa domain was fixed and the domain's original B-factors retained. This was performed due to the accuracy with which the structure of the 41kDa domain was known and due to the poor observation to parameter ratio of the 68kDa data, resulting from the incompleteness of data. After this round of PREPSTAGE the free R-factor was 41.6%. The corresponding side chains were now added and the Lego commands in O were used to obtain the most appropriate main chain and side chain rotamers, optimising their fit to the density. The 27kDa of the structure was then refined in PREPSTAGE and a new SIGMAA weighted map calculated together with the corresponding solvent flattened map.

Round 6 of refinement

Several more cycles of manual adjustment of the positions of side chains were carried out, with concomitant refinement in PREPSTAGE and calculation of new SIGMAA weighted maps. Unweighted $2F_o$-F_c maps generated using X-PLOR were also used at this stage, but did not reveal any obvious improvements that could be made. After the final cycle of refinement, the free R-factor was 26.3%. Conventional omit maps and simulated annealing omit maps were calculated, omitting residues in regions of interrupted or ambiguous density. However these gave little or no improvement over the previously calculated maps.

Round 7 of refinement

Fortuitously, shortly after the previous round of refinement a native data set which was 80.7% complete (>1σ) to 2.5Å became available. The present model for the 68kDa domain was refined at a resolution of 2.5Å in X-PLOR, with the linker domain removed and the 41kDa domain unfixed in the refinement. Due to the additional observations, the model was now also refined in SLOWCOOL and the free R-factor was 42.0% after completion of refinement. A new 2Fo-Fc map was generated with weighted Fourier coefficients in SIGMAA from the refined coordinates.

Round 8 of refinement

Careful inspection of the 2Fo-Fc electron density map confirmed that the connectivity of the linker domain and galactose binding domains were correct. It was however obvious that further manual rebuilding of certain regions, particularly the C-terminal end of the linker domain was required. This region was rebuilt, the coordinates for the linker domain added back and the model refined again in X-PLOR. After SLOWCOOL, the free R-factor for the model was 37.7%.

Round 9 of refinement

Simulated annealing omit maps were now calculated, omitting 15 residues at a time, starting at the end of the 41kDa domain and extending up to residue 536 of the galactose binding domain. Appropriate manual rebuilding using O was performed, but most of the modifications made were fairly subtle changes to the positions of side chains, or choice of side chain rotamers, or involved flipping

peptide bonds. Refinement in X-PLOR gave a free R-factor of 37.6% after SLOWCOOL. The availability of extra observations now allowed refinement of B-factors. The B-factors were refined as two separate groups for main chain and side chains and the free R-factor dropped to 37.2%.

Round 10 of refinement

Five more cycles of manual rebuilding with concomitant refinement in X-PLOR gave a model with a free R-factor of 35.4%. Simulated annealing omit maps were calculated after each round of refinement, omitting stretches of 15 residues over the residues with ϕ, φ angles in disallowed/generously allowed regions of the Ramachandran Plot. 2Fo-Fc maps were also calculated in X-PLOR after each round of refinement. In addition the location of a postulated sodium ion in the galactose binding domain was established (Chapter 9), the coordinates of which were now included in the model.

Round 11 of refinement

Individual B-factor refinement was implemented and a free R-factor of 34.0% was obtained.

Round 12 of refinement

The model was now refined against the complete data set with a 0σ cut-off. Firstly the individual B-factor refined model was refined in PREPSTAGE and then individual B-factors refined. This was repeated a second time to give a final model with a R-factor of 21.4%.

A summary of the refinement procedure is shown in Fig. 7.7 and Table 7.1 gives the refinement statistics.

Fig. 7.7 Refinement of the 68kDa form of *M.viridifaciens* neuraminidase.

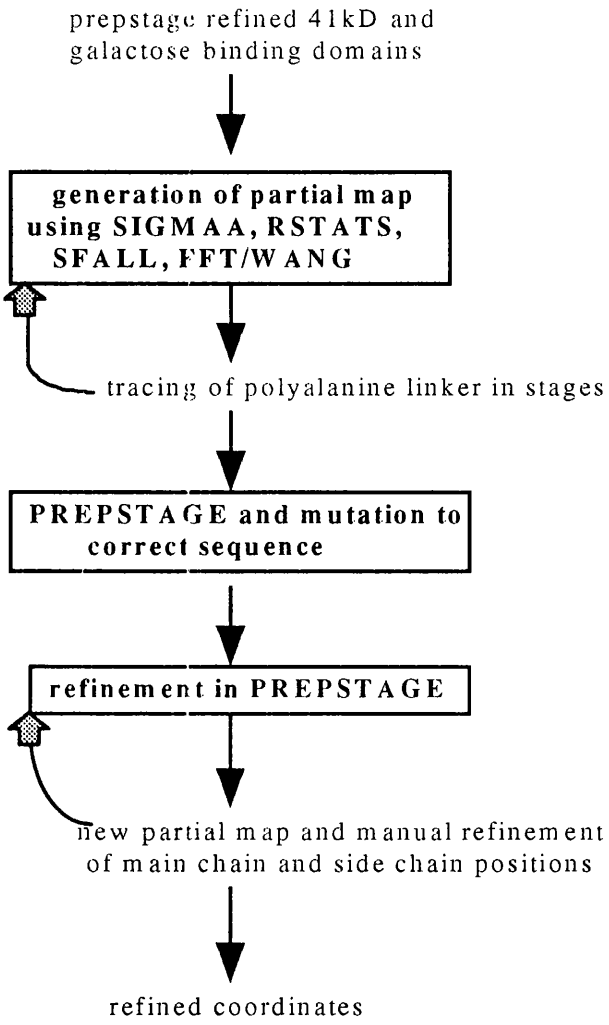


Table 7.1 Summary of the stages in the refinement of the 68kDa form of *M.viridifaciens* neuraminidase.

Round of refinement	stage of model	$R_{test\ set}$ (free R-factor) (%)	$R_{working\ set}$ (%)	resolution (Å)	fixed 41kDa domain?
1	rigid body refined 41kDa only	44.5	42.5	3.5	yes
2	rigid body refined 41kDa & galactose binding domain	47.1	43.7	3.5	yes
3	prepstage without linker domain	44.1	34.5	3.5	yes
4 & 5	prepstage with polyalanine linker domain	41.6	30.8	3.5	yes
6	prepstage with complete linker domain	29.7	24.9	3.5	yes
7	slowcool without linker domain against new data	42.3	30.7	2.5	no
8	slowcool with linker domain after manual rebuilding	37.7	26.1	2.5	no
9	slowcool and B-group refinement following manual rebuilding	37.2	23.6	2.5	no
10	5 cycles of manual rebuilding with concomitant refinement in X-PLOR. Coordinates of sodium ion included in model	35.4	22.3	2.5	no
11	individual B-factor refinement	34.0	21.7	2.5	no
12	refinement against complete data set (2 cycles of prepstage and individual B-factor refinement)	-	R_{cryst} = 21.4%	2.5	no

Discussion and quality of the final model

Since a data set was initially only available to a medium resolution of 3.5Å, with a data completeness of 57% (>1σ) the refinement had to be approached with caution. However with the availability of the 2.5Å data set, the refinement progressed well and confirmed that the connectivity of the linker domain was correct. The crystallographic R-factor of the refined model is 21.4% for all reflections between 6 and 2.5Å and the model includes residues 47 to 647 and one sodium ion, but no water molecules. Table 7.2 summarizes the refinement statistics for the final model.

Table 7.2 Refinement statistics for the final model of the 68kDa form of *M.viridifaciens* neuraminidase.

number of non-hydrogen protein atoms	4536
number of water molecules	0
resolution range	6.0 - 2.5Å
number of reflections	18,632
R-factor for all data	21.4%
rms deviation in: bond lengths bond angles	 0.013Å 2.066°
Average temperature factors: main chain side chain	 17.68Å^2 18.97Å^2

The same tests were applied to judge the quality of the refined model as were used with the 41kDa *M.viridifaciens* neuraminidase structure (Chapter 5 Quality of the final model). The 3D-1D score was greater than zero for all regions of the structure, indicating that all side chains were in favourable environments

(fig. 7.8). The Luzatti plot for the refined structure gave a mean positional error of between 0.25 and 0.30Å for all data between 6 and 2.5Å, with a 2σ cut-off (fig. 7.9). Fig. 8.0 shows the Ramachandran plot for the final model produced using Procheck. There were 2 residues with ϕ, φ angles in disallowed regions of the Ramachandran plot and 6 residues with ϕ, φ angles in generously allowed regions.

Fig. 7.8 The 3D-1D score for the final refined 68kDa model.

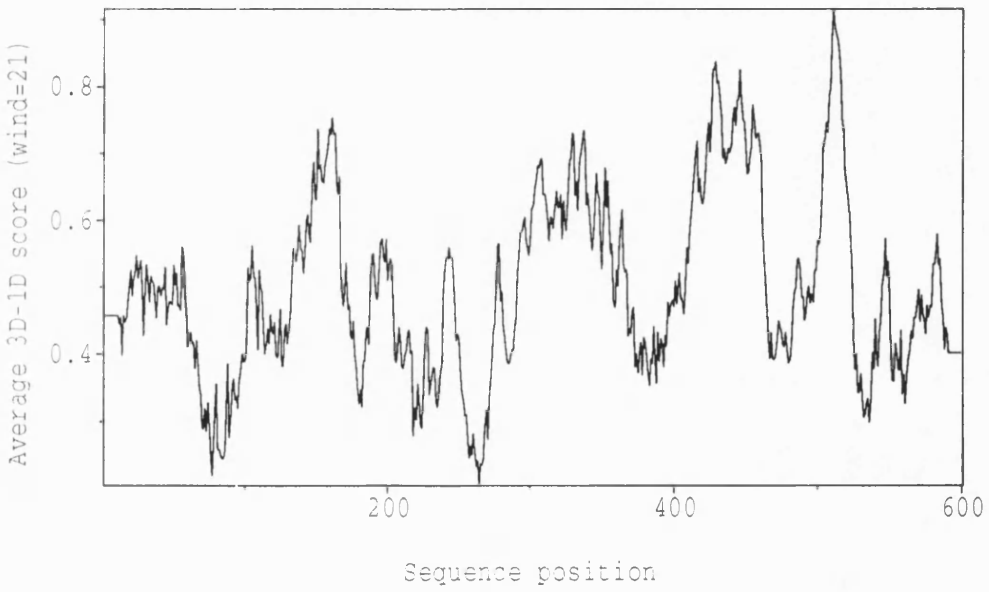


Fig. 7.9 Luzatti plot for the refined 68kDa model. The bold line gives the observed R-factor. The broken lines represent the theoretical values associated with a given positional error in angstroems.

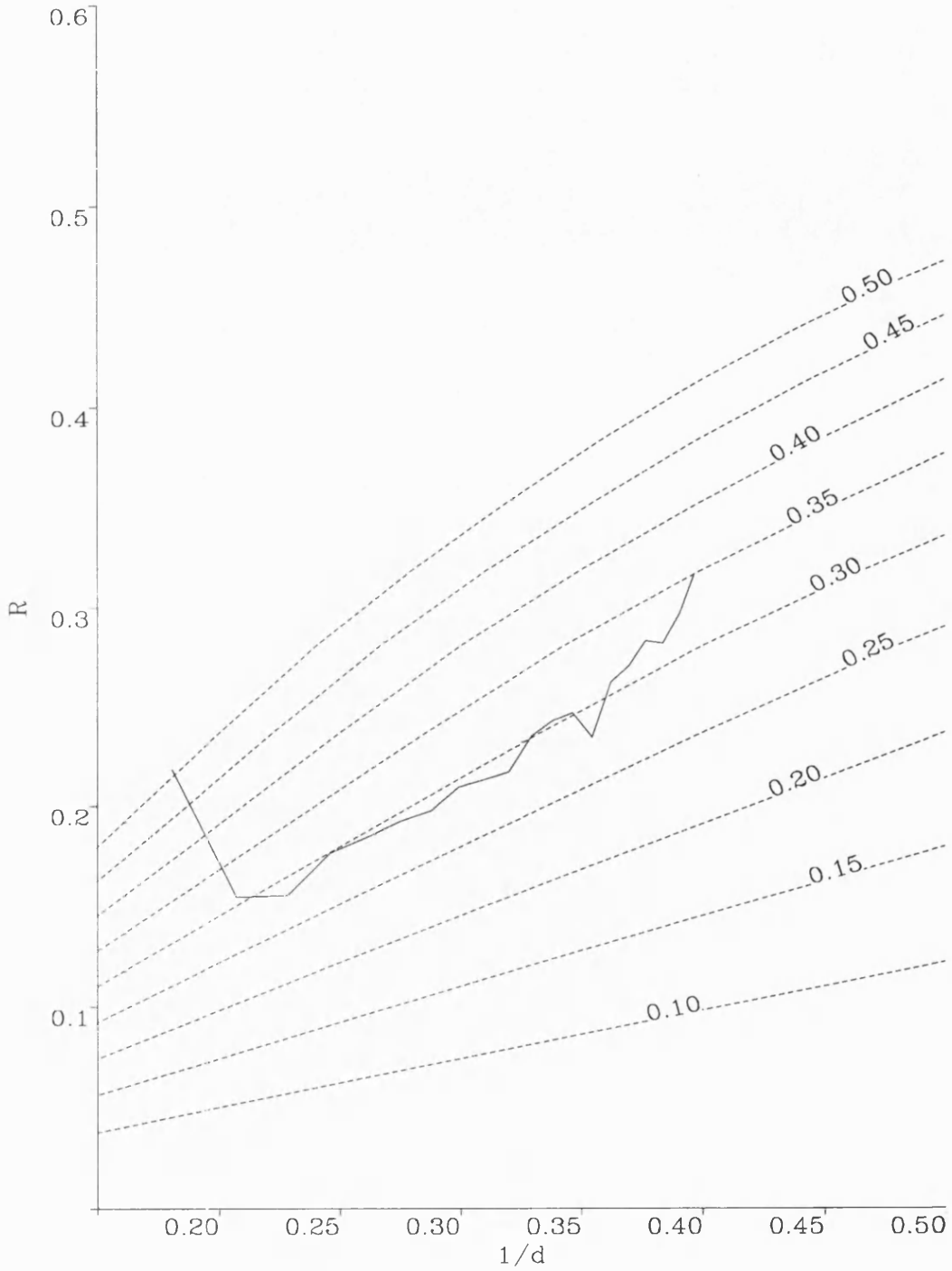
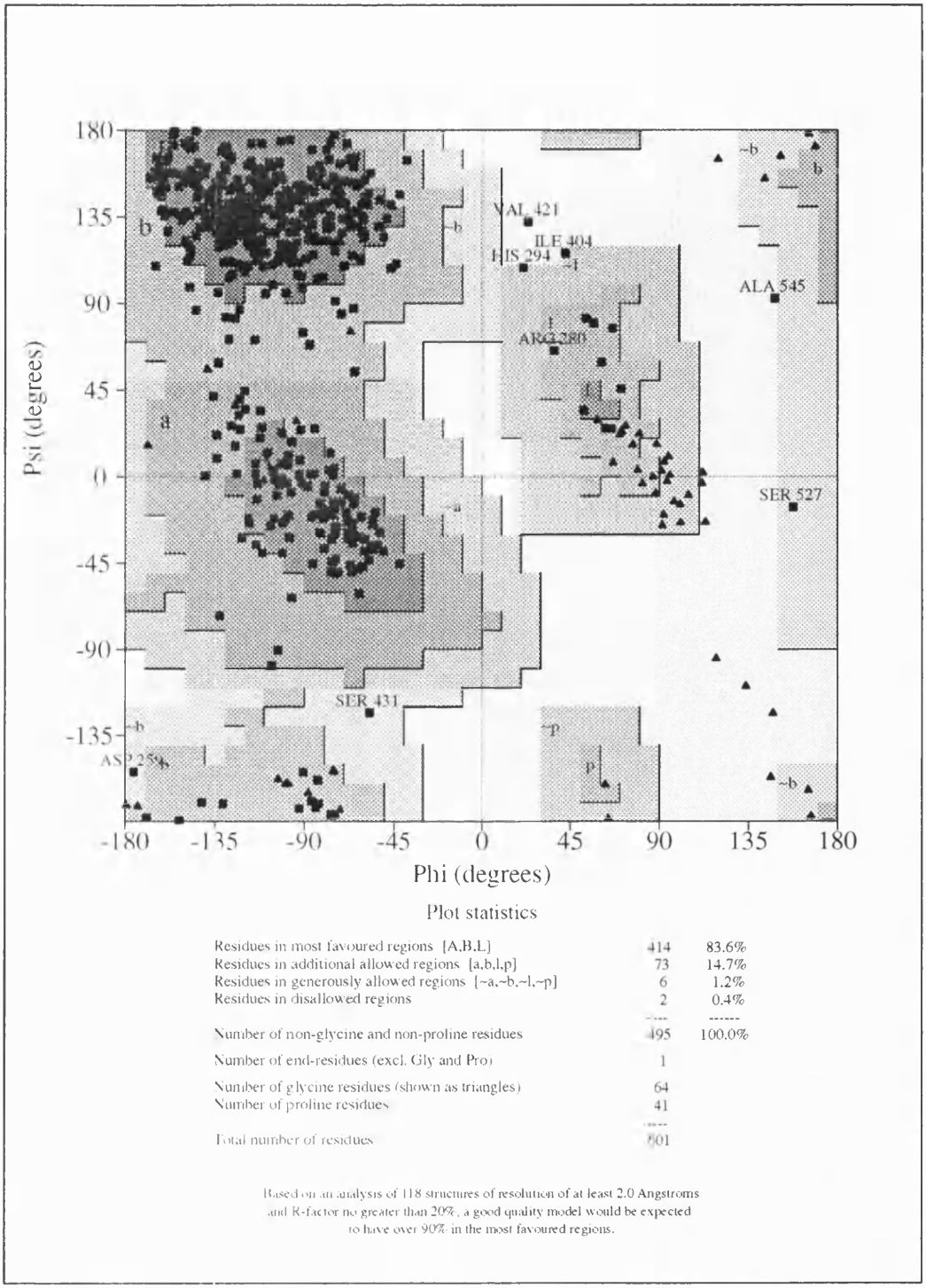


Fig. 8.0 The Ramachandran plot for the refined model of 68kDa *M.viridifaciens* neuraminidase.



Plot statistics

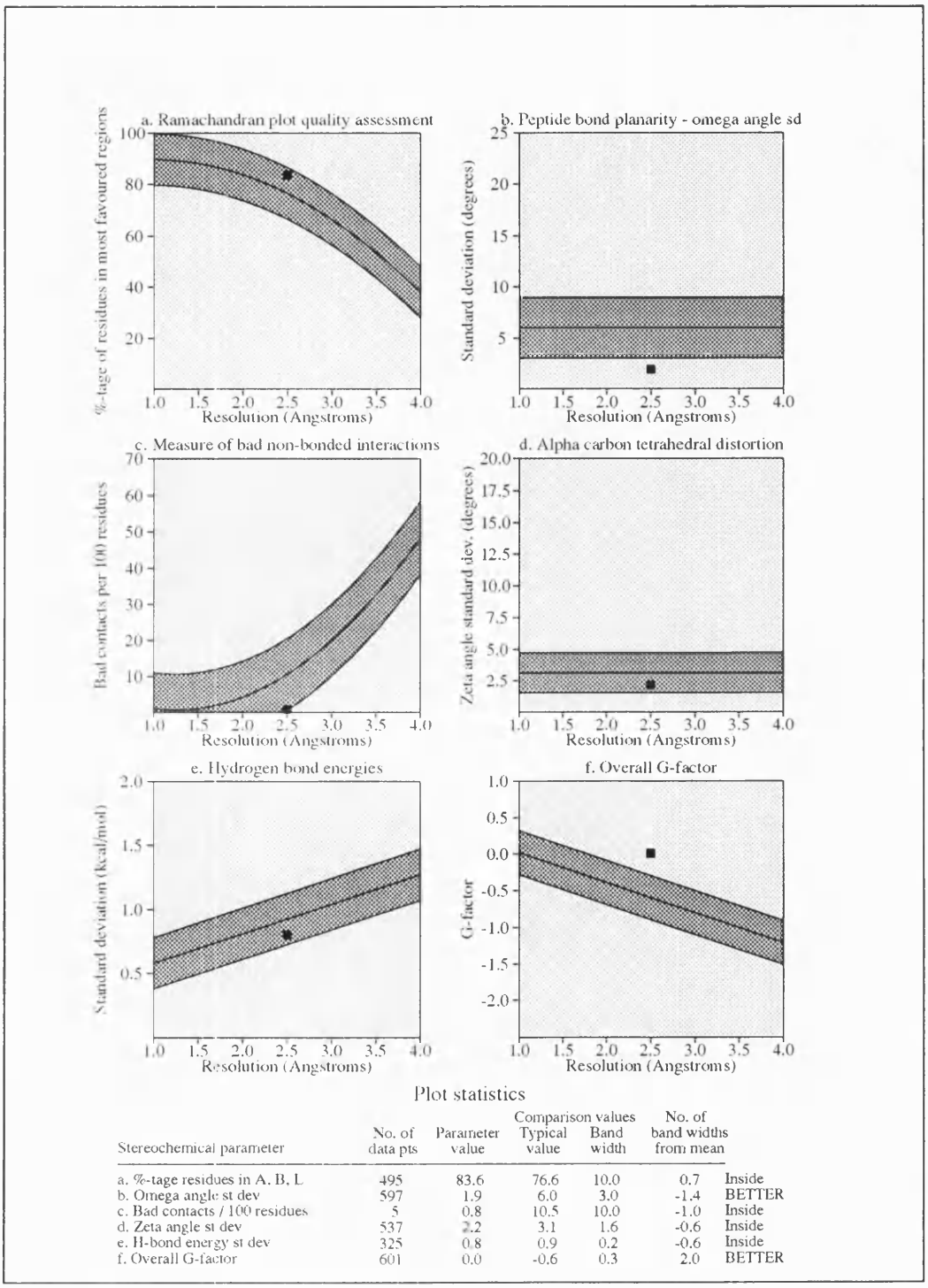
Residues in most favoured regions [A,B,L]	414	83.6%
Residues in additional allowed regions [a,b,l,p]	73	14.7%
Residues in generously allowed regions [~a,~b,~l,~p]	6	1.2%
Residues in disallowed regions	2	0.4%
Number of non-glycine and non-proline residues	495	100.0%
Number of end-residues (excl. Gly and Pro)	1	
Number of glycine residues (shown as triangles)	64	
Number of proline residues	41	
Total number of residues	601	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

The various main chain and side chain parameters obtained from Procheck for the refined model are illustrated as a function of residue number in fig. 8.1. The average temperature factors for main chain and side chain atoms were 17.68Å^2 and 18.97Å^2 respectively and the temperature factors for the refined model as a function of the residue number are given in fig. 8.2. The rms deviation in bond lengths and angles from ideality were 0.013Å and 2.066° respectively for the final refined model.

Fig. 8.1 (**a**) Main chain and (**b**) side chain parameters for the refined model of the 68kDa form of *M.viridifaciens* neuraminidase.

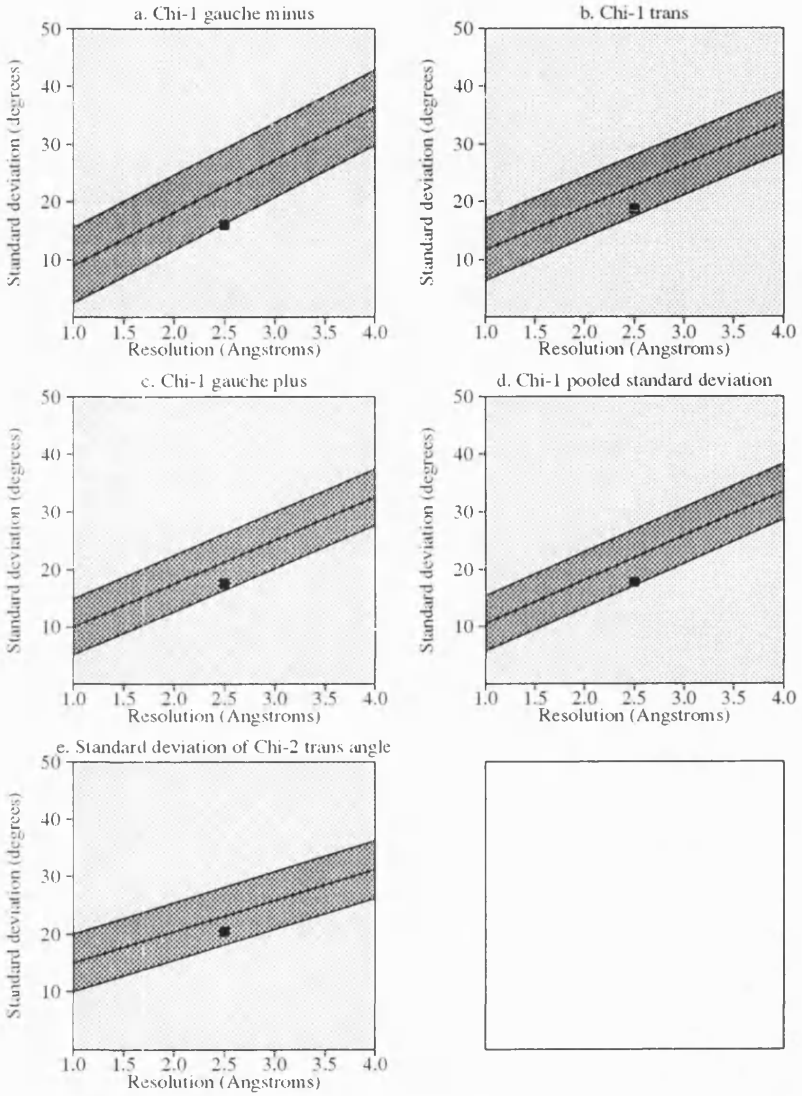
(a)



Plot statistics

Stereochemical parameter	No. of data pts	Parameter value	Comparison values Typical value	Band width	No. of band widths from mean	
a. %-tage residues in A, B, L	495	83.6	76.6	10.0	0.7	Inside
b. Omega angle st dev	597	1.9	6.0	3.0	-1.4	BETTER
c. Bad contacts / 100 residues	5	0.8	10.5	10.0	-1.0	Inside
d. Zeta angle st dev	537	2.2	3.1	1.6	-0.6	Inside
e. H-bond energy st dev	325	0.8	0.9	0.2	-0.6	Inside
f. Overall G-factor	601	0.0	-0.6	0.3	2.0	BETTER

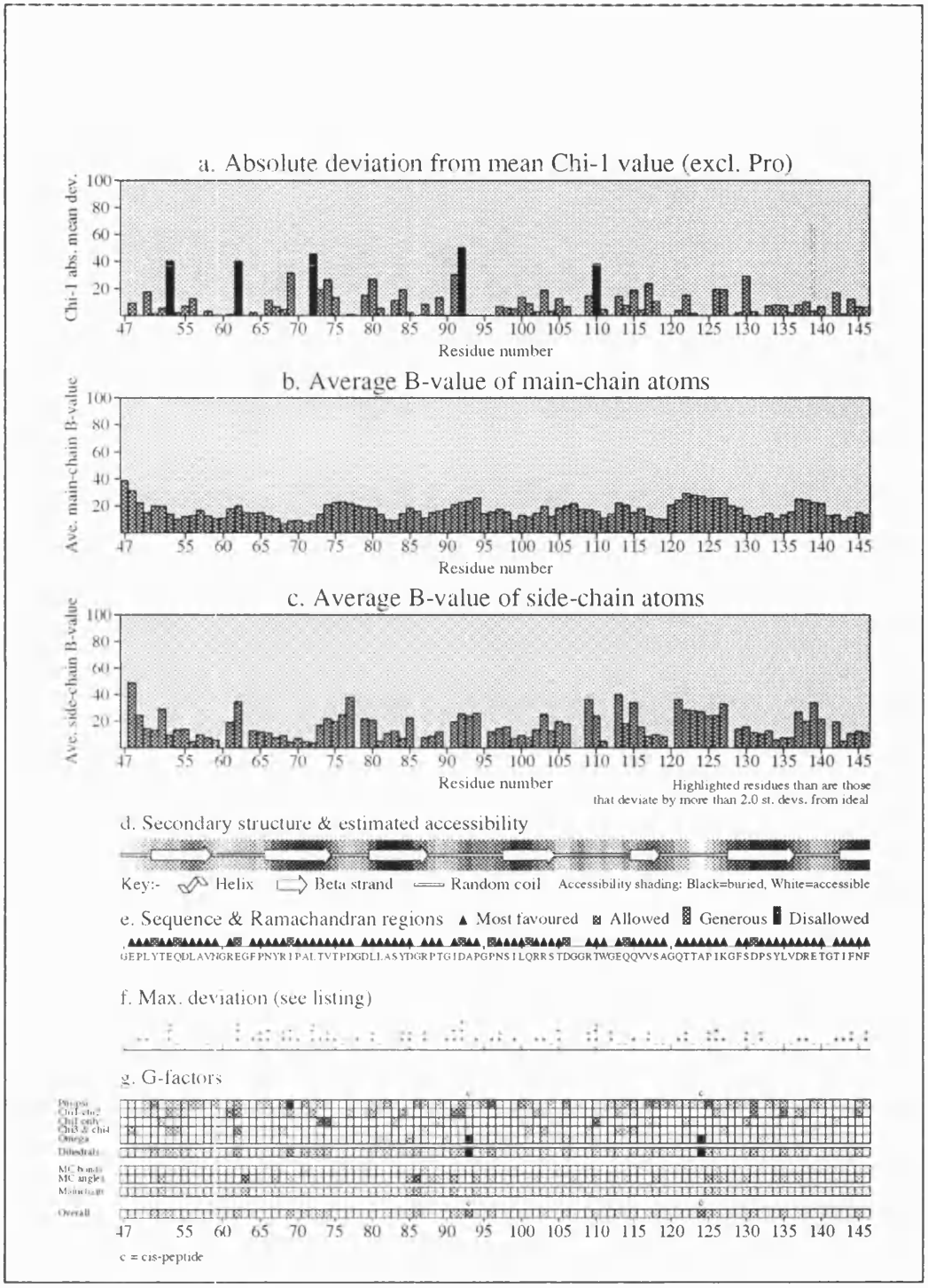
(b)

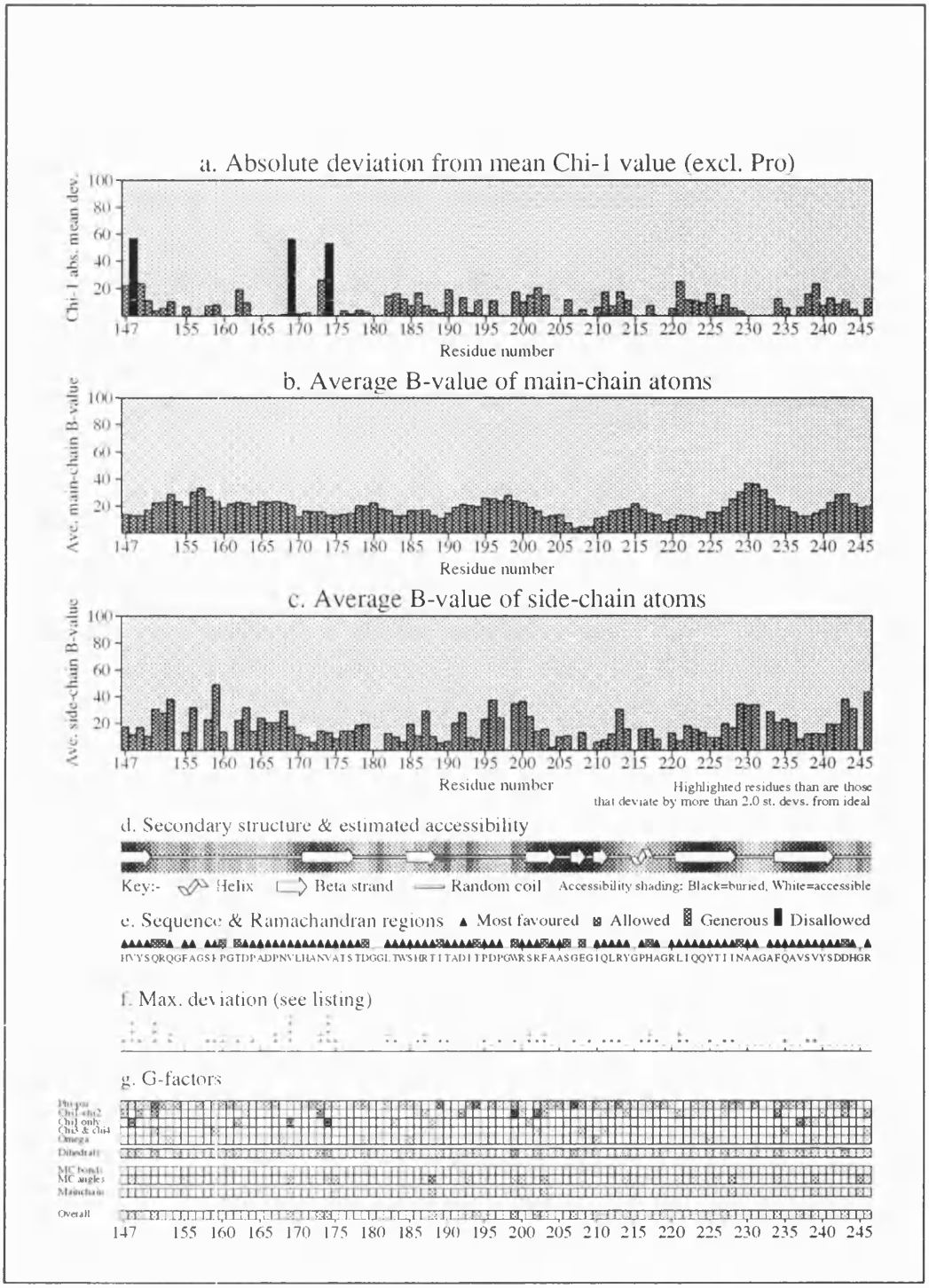


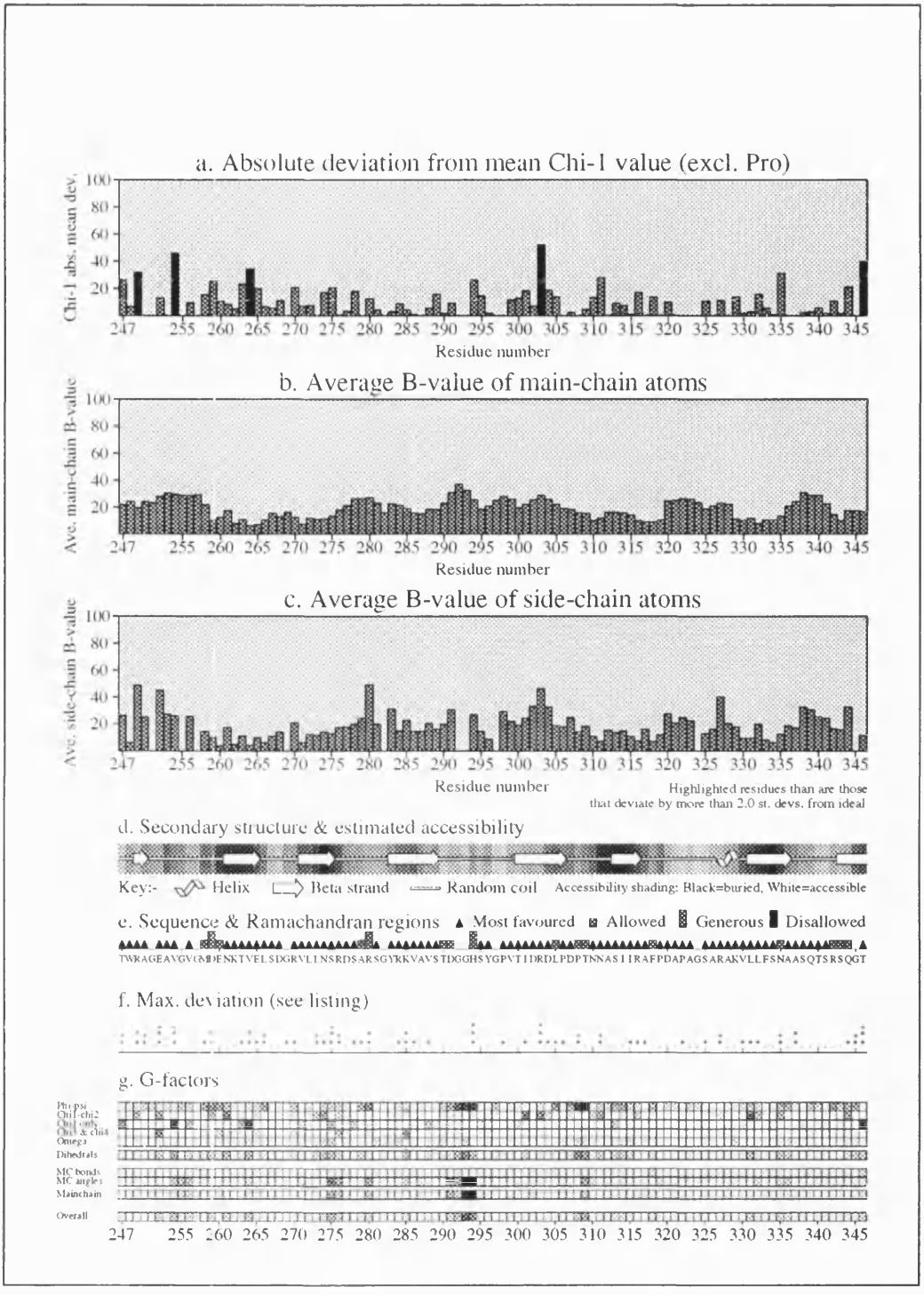
Plot statistics

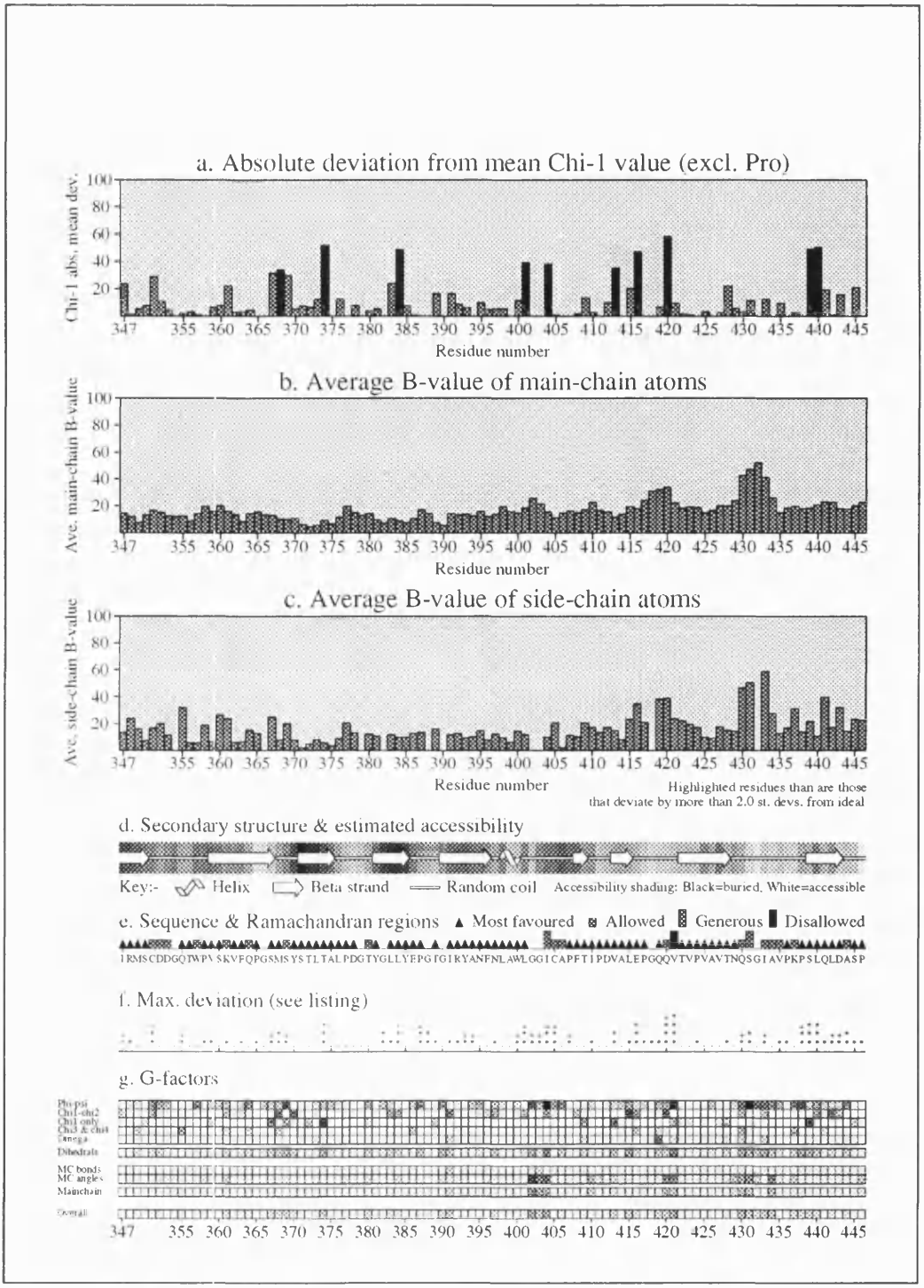
Stereochemical parameter	No. of data pts	Parameter value	Comparison values Typical value	Band width	No. of band widths from mean	
a. Chi-1 gauche minus st dev	105	16.0	22.7	6.5	-1.0	BETTER
b. Chi-1 trans st dev	124	18.8	22.7	5.3	-0.7	Inside
c. Chi-1 gauche plus st dev	207	17.6	21.3	4.9	-0.8	Inside
d. Chi-1 pooled st dev	436	17.8	22.0	4.8	-0.9	Inside
e. Chi-2 trans st dev	107	20.5	23.1	5.0	-0.5	Inside

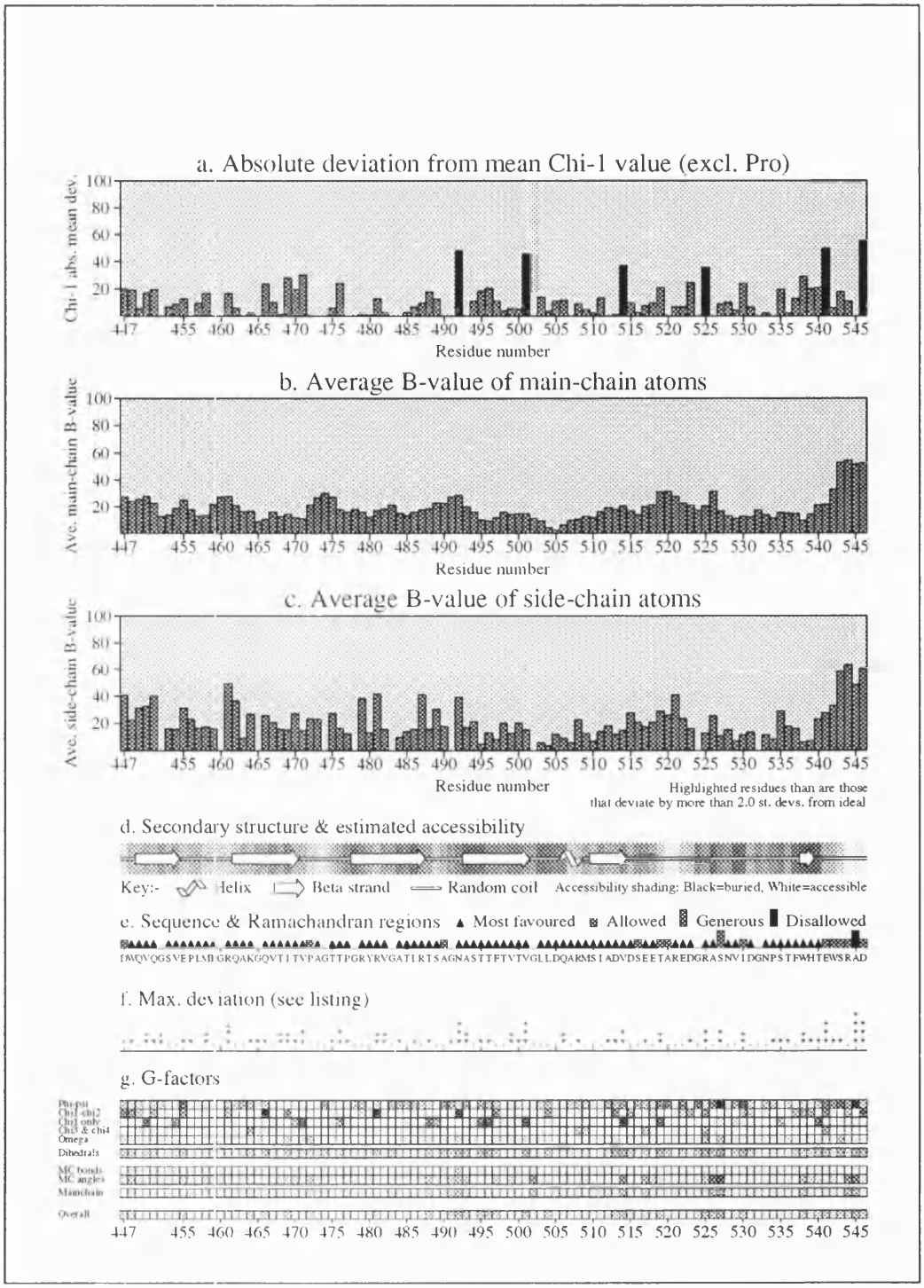
Fig. 8.2 Temperature factors as a function of residue for the refined model of 68kDa *M.viridifaciens* neuraminidase.

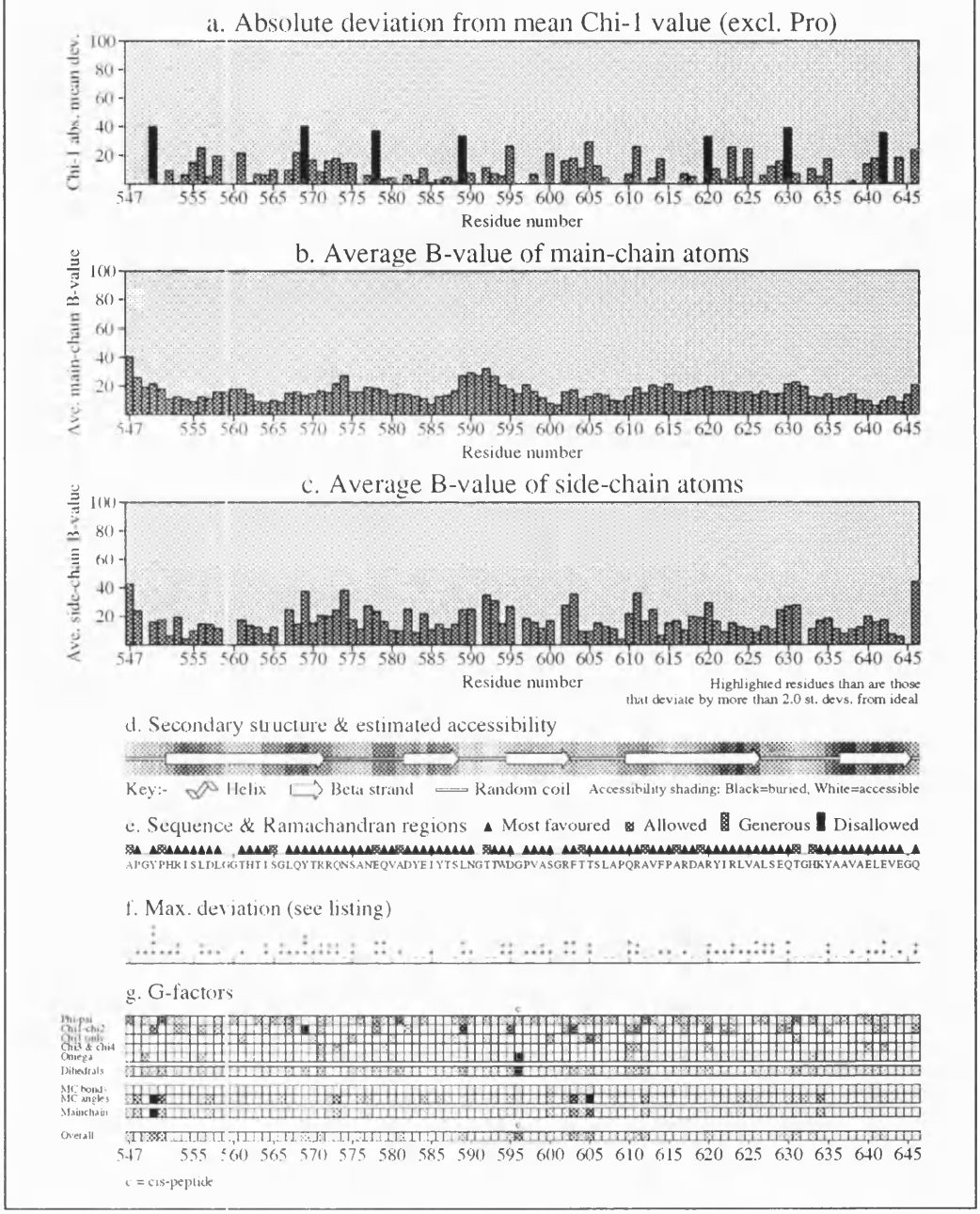


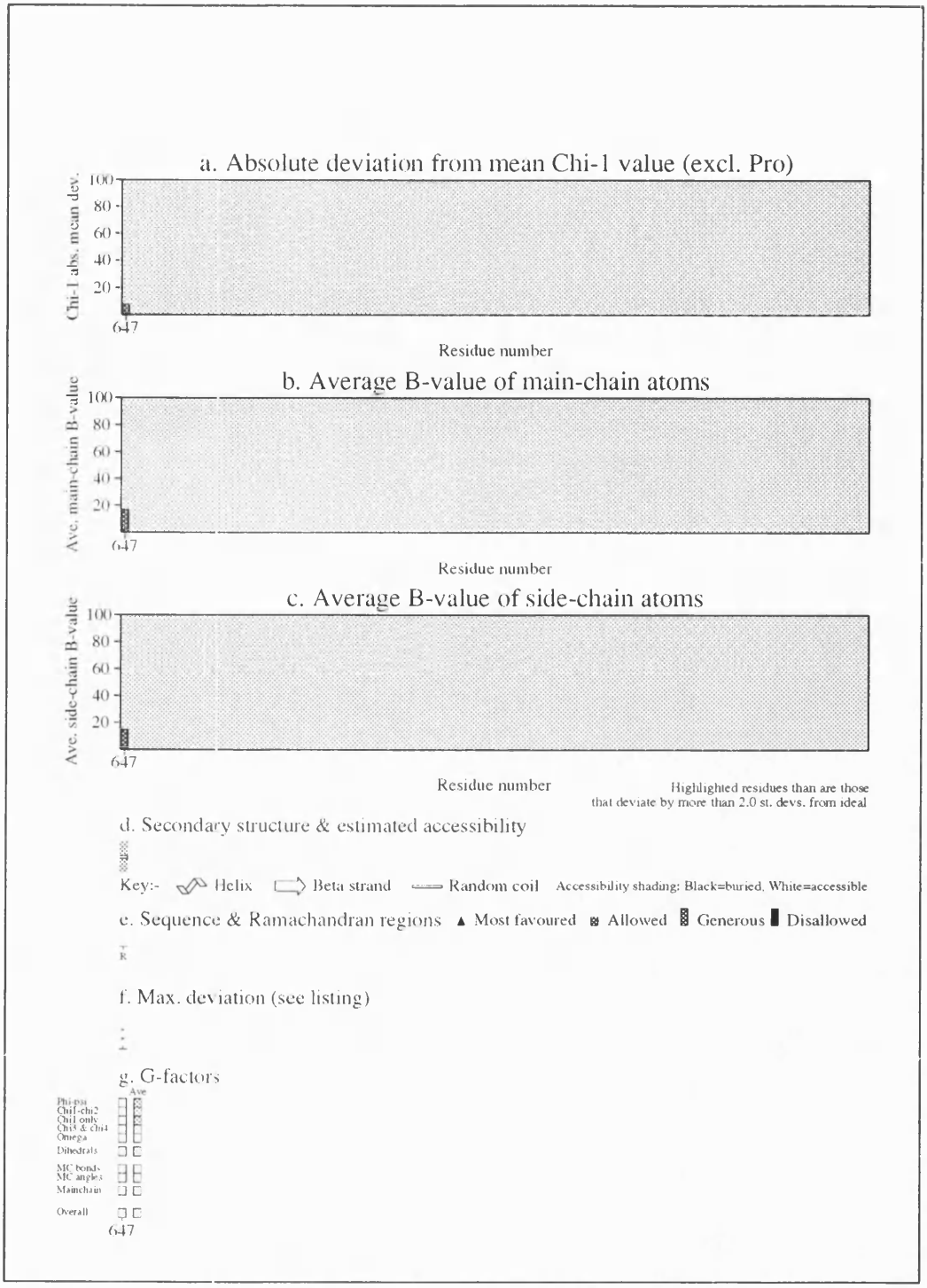

a. Absolute deviation from mean Chi-1 value (excl. Pro)
Chi-1 abs. mean dev.
Residue number
b. Average B-value of main-chain atoms
Ave. main-chain B-value
Residue number
c. Average B-value of side-chain atoms
Ave. side-chain B-value
Residue number
Highlighted residues than are those that deviate by more than 2.0 st. devs. from ideal
d. Secondary structure & estimated accessibility
Key:- Helix Beta strand Random coil Accessibility shading: Black=buried, White=accessible
e. Sequence & Ramachandran regions Most favoured Allowed Generous Disallowed
f. Max. deviation (see listing)
g. G-factors
Overall


a. Absolute deviation from mean Chi-1 value (excl. Pro)
Chi-1 abs. mean dev.
Residue number
b. Average B-value of main-chain atoms
Ave. main-chain B-value
Residue number
c. Average B-value of side-chain atoms
Ave. side-chain B-value
Residue number
Highlighted residues than are those that deviate by more than 2.0 st. devs. from ideal
d. Secondary structure & estimated accessibility
Key:- Helix Beta strand Random coil Accessibility shading: Black=buried, White=accessible
e. Sequence & Ramachandran regions Most favoured Allowed Generous Disallowed
f. Max. deviation (see listing)
g. G-factors
Phi-psi
Chi1-chi2
Chi1-omega
Chi1 & chi2
Omega
Dihedrals
MC bonds
MC angles
Mainchain
Overall


a. Absolute deviation from mean Chi-1 value (excl. Pro)
Chi-1 abs. mean dev.
Residue number
b. Average B-value of main-chain atoms
Ave. main-chain B-value
Residue number
c. Average B-value of side-chain atoms
Ave. side-chain B-value
Residue number
Highlighted residues than are those that deviate by more than 2.0 st. devs. from ideal
d. Secondary structure & estimated accessibility
Key:- Helix Beta strand Random coil Accessibility shading: Black=buried, White=accessible
e. Sequence & Ramachandran regions Most favoured Allowed Generous Disallowed
f. Max. deviation (see listing)
g. G-factors
Dihedrals
MC bonds
MC angles
Mainchain
Overall


a. Absolute deviation from mean Chi-1 value (excl. Pro)
Chi-1 abs. mean dev.
Residue number
b. Average B-value of main-chain atoms
Ave. main-chain B-value
Residue number
c. Average B-value of side-chain atoms
Ave. side-chain B-value
Residue number
Highlighted residues than are those that deviate by more than 2.0 st. devs. from ideal
d. Secondary structure & estimated accessibility
Key:- Helix Beta strand Random coil Accessibility shading: Black=buried, White=accessible
e. Sequence & Ramachandran regions Most favoured Allowed Generous Disallowed
f. Max. deviation (see listing)
g. G-factors
Phi-psi
Chi1-chi2
Chi1 only
Chi3 & chi4
Omega
Dihedrals
MC bonds
MC angles
Mainchain
Overall
447 455 460 465 470 475 480 485 490 495 500 505 510 515 520 525 530 535 540 545


a. Absolute deviation from mean Chi-1 value (excl. Pro)
Chi-1 abs. mean dev.
Residue number
b. Average B-value of main-chain atoms
Ave. main-chain B-value
Residue number
c. Average B-value of side-chain atoms
Ave. side-chain B-value
Residue number
Highlighted residues than are those that deviate by more than 2.0 st. devs. from ideal
d. Secondary structure & estimated accessibility
Key:- Helix Beta strand Random coil Accessibility shading: Black=buried, White=accessible
e. Sequence & Ramachandran regions Most favoured Allowed Generous Disallowed
APGYPHRISLDLGGTHTISGLQYTRKQNSANEQVADYEIYTSLNGTTWDGPVASGREFTTSLAPQRAVFPARDARYIRLVALSEQTGHKYAAVAELEVEGQ
f. Max. deviation (see listing)
g. G-factors
Phi-psi
Chi1-chi2
Chi1-only
Chi3 & chi4
Omega
Dihedrals
MC bond
MC angles
Mainchain
Overall
c = cis-peptide


a. Absolute deviation from mean Chi-1 value (excl. Pro)
Chi-1 abs. mean dev.
100
80
60
40
20
647
Residue number
b. Average B-value of main-chain atoms
Ave. main-chain B-value
100
80
60
40
20
647
Residue number
c. Average B-value of side-chain atoms
Ave. side-chain B-value
100
80
60
40
20
647
Residue number
Highlighted residues than are those
that deviate by more than 2.0 st. devs. from ideal
d. Secondary structure & estimated accessibility
Key:- Helix Beta strand Random coil Accessibility shading: Black=buried, White=accessible
e. Sequence & Ramachandran regions ▲ Most favoured Allowed Generous Disallowed
f. Max. deviation (see listing)
g. G-factors
Ave
Phi-psi
Chi1-chi2
Chi1 only
Chi3 & chi4
Omega
Dihedrals
MC bonds
MC angles
Mainchain
Overall
647

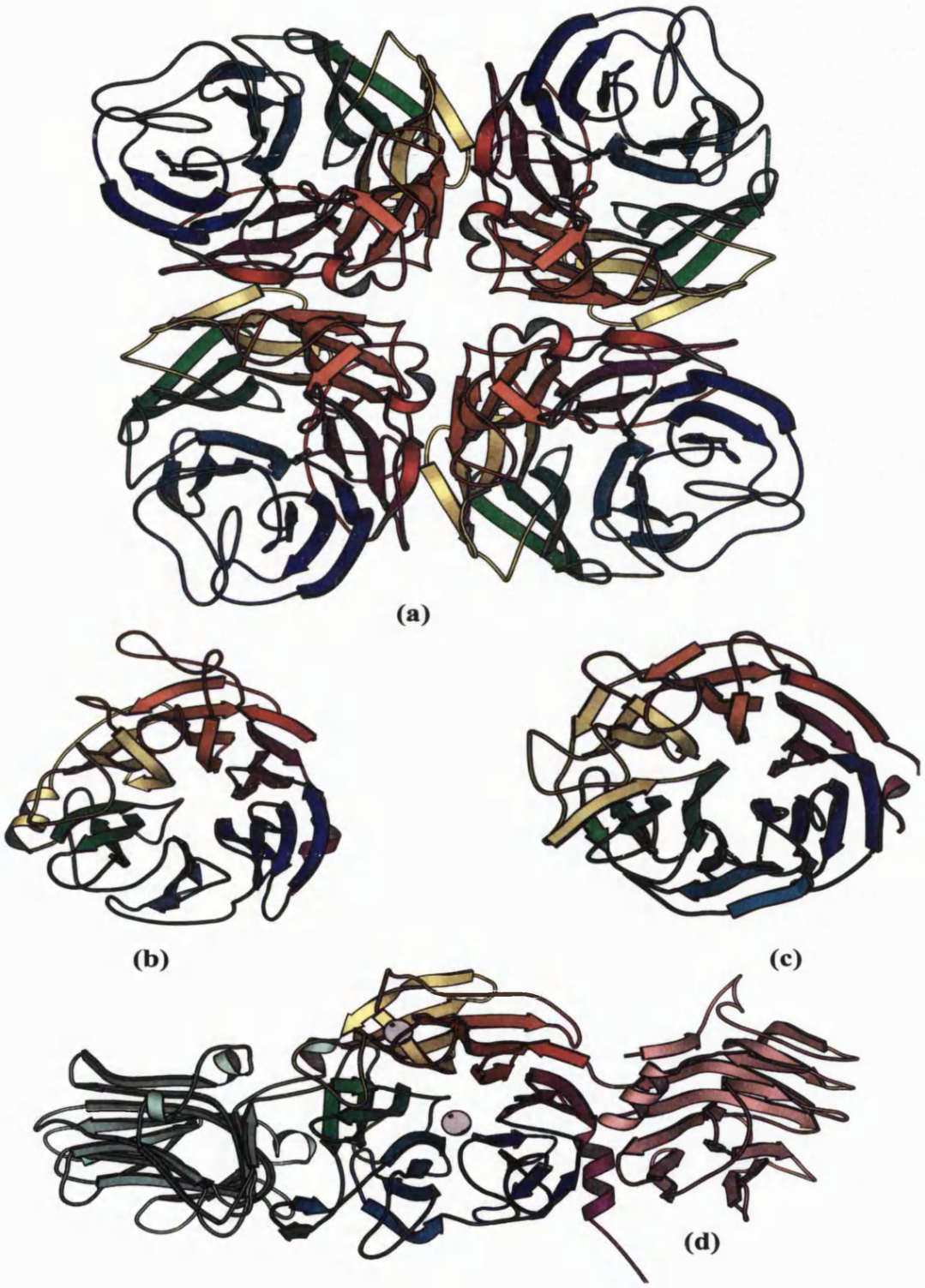
Chapter 8

THE 41kDa *M.viridifaciens* NEURAMINIDASE STRUCTURE

Protein architecture

The 41kDa *M.viridifaciens* neuraminidase is monomeric and exhibits the typical neuraminidase β-propeller fold. This consists of six four stranded antiparallel β-pleated sheets arranged on the blades of a propeller. In addition, there is one short 4 residue α-helix at the C-terminus of the enzyme and two short 3_{10} helices, as defined in the DSSP (Kabsch & Sander, 1983) definition of secondary structure. One of the 3_{10} helices is located between residues 190 and 193 and the other is located between residues 327 and 329. A view down the six-fold axis of the monomer is shown in fig. 8.1, together with the same view down the six-fold axes of the neuraminidases from influenza N8, *S.typhimurium* and *V.cholerae*, for comparison.

Fig. 8.1 A view down the six-fold axis of three bacterial neuraminidases and a view down the four-fold axis of one viral neuraminidase. (a) Influenza N8 neuraminidase (b) *S.typhimurium* neuraminidase (c) *M.viridifaciens* neuraminidase and (d) *V.cholerae* neuraminidase. Note the lectin-like domains on either side of the central catalytic domain of *V.cholerae* neuramidase.



Comparison with other neuraminidase structures

In spite of possessing a low sequence identity with other neuraminidases (table 8.1), the 41kDa *M.viridifaciens* neuraminidase shows a high degree of structural homology. The 41kDa form of *M.viridifaciens* neuraminidase has the highest structural homology with *V.cholerae* neuraminidase. The two neuraminidases share 26% sequence identity, yet superposition of the two structures results in a rms deviation of 2.31Å over 336 Cα atoms. Influenza neuraminidase of subtype N9 has the lowest structural similarity with the 41kDa form of *M.viridifaciens* neuraminidase; superposition of the two structures gives a rms deviation of 3.69Å over 296 Cα atoms. However Influenza N8 and N2 subtypes produce similar rms deviations with the 41kDa form of *M.viridifaciens* neuraminidase. Overall structural homologies between the 41kDa *M.viridifaciens* neuraminidase and all other neuraminidase structures known to date, are summarized in table 8.2. Rms deviations were generated using the program MNYFIT from the COMPOSER suite of modelling software(Sutcliffe *et al.*, 1987). MNYFIT performs the superposition of two or more protein structures and determines the location of structurally equivalent residues, using this information to define the family framework. Six topologically equivalent active site residue atoms were chosen for the initial superposition by MNYFIT. These are listed in table 8.3.

Table 8.1 Sequence identity matrix for viral and non-viral neuraminidases.

%	mhc g9	T.cruzi	*V.cholerae*	st[2]	N8	N9	N2	B/Beijing	B/Lee
mv[1]	33.44	21.77	25.92	19.22	16.23	15.39	16.19	18.30	19.77

[1] *M.viridifaciens* neuraminidase
[2] *S.typhimurium* neuraminidase

Table 8.2 rmsds (Å) between the 41kDa *M.viridifaciens* neuraminidase and the other neuraminidase structures. The numbers in brackets are the number of equivalences found by COMPOSER and used in the superposition.

Å	*V.cholerae*	st[2]	N8	N9	N2	B/Beijing	B/Lee
mv[1]	2.31 (336)	2.59 (335)	3.57 (295)	3.69 (296)	3.65 (294)	3.23 (302)	3.20 (302)

[1] *M.viridifaciens* neuraminidase

[2] *S.typhimurium* neuraminidase

Table 8.3 Topologically equivalent active site residues used in the initial superposition by COMPOSER.

mv[1]	*V.cholerae*	st[2]	N8	N9	N2	B/Beijing	B/Lee
Arg 68	Arg 224	Arg 37	Arg 118	Arg 118	Arg 118	Arg A115	Arg 116
Arg 342	Arg 712	Arg 309	Arg 371	Arg 371	Arg 371	Arg A373	Arg 374
Arg 276	Arg 635	Arg 246	Arg 292	Arg 292	Arg 292	Arg A291	Arg 292
Tyr 370	Tyr 740	Tyr 342	Tyr 406	Tyr 406	Tyr 406	Tyr A408	Tyr 409
Glu 386	Glu 756	Glu 361	Glu 425	Glu 425	Glu 425	Glu A427	Glu 428
Glu 260	Glu 619	Glu 231	Glu 277	Glu 277	Glu 277	Glu A275	Glu 276

[1] *M.viridifaciens* neuraminidase
[2] *S.typhimurium* neuraminidase

Conserved sequence motifs

In spite of the low sequence identity observed between bacterial neuraminidases (Chapter 1 Introduction), there remain two conserved sequence motifs found across bacterial neuraminidase sequences.

1. Asp-boxes

The 41kDa form of *M.viridifaciens* neuraminidase contains 5 so called Asp-boxes. Asp-boxes are present in all bacterial neuraminidases characterized to date and have the consensus sequence Ser/Thr-X-Asp-[X]-Gly-X-Thr-Trp/Phe. The five Asp-boxes occur in the same topologically equivalent positions as observed for the other two bacterial neuraminidases (*V.cholerae* and *S.typhimurium*): at the turns between the third and fourth strands of each sheet. The Asp-boxes are once again distant from the active site and have no role in catalysis. The aromatic side chains of Asp-boxes tuck neatly in between adjacent strands and the aspartic acid side chains point out into solvent. The five Asp-boxes of *M.viridifaciens* neuraminidase were superimposed with respect to the first Asp-box using Lsq_explicit and Lsq_improve in O and table 8.4 summarizes the results. Lsq_explicit performs a least squares comparison of two structures via explicit atoms. Lsq_improve improves on a given transformation produced by Lsq_explicit. Lsq_improve employs an algorithm involving a search for structure fragments that can be aligned within a given cut off limit and tries to improve the transformation and the matched atoms. The algorithm cycles until either the same set of atoms are selected or until 10 cycles are made.

Table 8.4 Asp-box similarity in the 41kDa form of *M.viridifaciens* neuraminidase. Values are given as rmsds (Å) between Asp-box 1 (D1) and Asp-boxes 2 (D2)-5 (D5). Sequences of the Asp-boxes are also given.

rmsd (Å)	D1	D2	D3	D4	D5
D1	0.00	0.18	0.32	1.87	0.19

D1: S T D G G R T W

D2: S T D G G L T W

D3: S D D H G R T W

D4: S T D G G H S Y

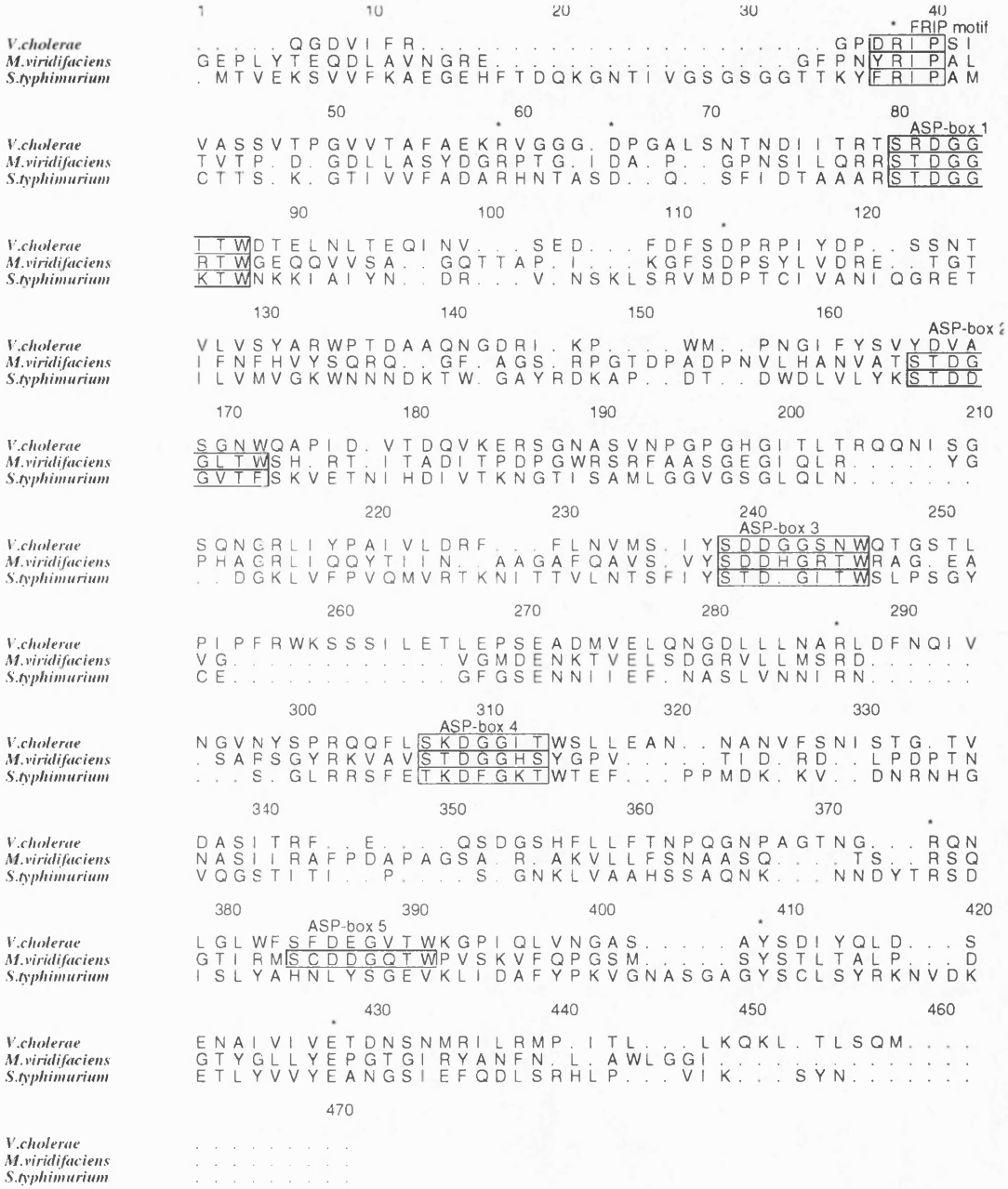
D5: S C D D G Q T W

It is apparent from table 8.4 that the outlier of the five Asp-boxes is Asp-box 4, which has a rmsd of 1.87Å with Asp-box 1. In Asp-box 4 a tyrosine replaces the tryptophan which is conserved in the other four Asp-boxes. The OH of the tyrosine hydrogen bonds to the ND1 atom of a histidine two residues prior to the tyrosine in the sequence. This additional hydrogen bonding, not present in the other Asp-boxes, results in a distortion of the loop from its normal conformation. Since the catalytic activity of the enzyme is not affected, the Asp-box motif appears to merely dictate the fold and integrity of the propeller and may not be involved in a secretory mechanism.

2. The FRIP motif

The sequence Phe-Arg-Ile-Pro, or small variations on this theme, occurs at the N-terminal end of all bacterial neuraminidase sequences. In the case of the neuraminidase from *M.viridifaciens*, the Phe is replaced by Tyr. The conserved Arginine residue in the FRIP motif represents one of the catalytic triad of arginines in the active site, which stabilizes the carboxylate group of sialic acid. A sequence alignment of the 41kDa form of *M.viridifaciens* neuraminidase against the sequence of the neuraminidase from *S.typhimurium* and *V.cholerae* is shown in fig. 8.2. The FRIP and Asp-box motifs are highlighted.

Fig. 8.2 Sequence alignment of three bacterial neuraminidase sequences: 41kDa *M.viridifaciens*, *S.typhimurium* and *V.cholerae* neuraminidase (neuraminidase domain only). Conserved catalytic residues are marked with an asterix and the FRIP and Asp-box motifs are boxed. Note the absence of Asp-box 2 in *V.cholerae* neuraminidase.



The Active site - obtaining the inhibitor complex

The position of the active site in the 41kDa form of *M.viridifaciens* neuraminindase was confirmed by soaking a solution of the inhibitor 2-deoxy-2,3-dehydro-N-acetylneuraminic acid (DANA) (fig. 1.3, Chapter 1) at a concentration of 5mM into a 41kDa crystal for 3 hours and collecting inhibitor complex data.

Data collection and analysis

All inhibitor data were collected from the in-house rotating anode source and a data set which was 83.0% complete (>1σ) to 2Å was obtained by merging two separate data sets in XSCALE. All data processing was performed using the XDS suite of software. The data collection statistics are given in table 8.5.

Parameters for data collection were as follows:

crystal to detector distance = 10cm

X-ray generator settings = 50kV, 80mA

exposure time = 120 - 180 seconds/frame

2θ swing angle = 20°

temperature ≈ 21°C

Table 8.5 Data collection statistics for the 41kDa inhibitor complex.

resolution (Å)	no. of observations	no. of unique reflections	completeness (>1σ) %	completeness (>3σ) %	R_I (%)[a]
∞ - 10.0	1169	230	99.1	96.14	4.9
10.0 - 5.0	8403	1413	99.1	95.8	4.5
5.0 - 4.0	8832	1467	99.3	97.3	4.8
4.0 - 3.0	22549	4074	98.7	94.6	5.8
3.0 - 2.5	23324	4963	95.2	85.7	8.3
2.5 - 2.0	28763	8219	67.4	52.7	11.2
∞ - 2.0	93043	20366	83.0	72.8	6.3

cell dimensions	a	b	c	α	β	γ
	48.00	82.86	84.94	90.00	90.00	90.00

$$^{a}R_I = \frac{\sum_h \sum_i |(I_h - I_{hi})|}{\sum_h \sum_i I_{hi}}$$

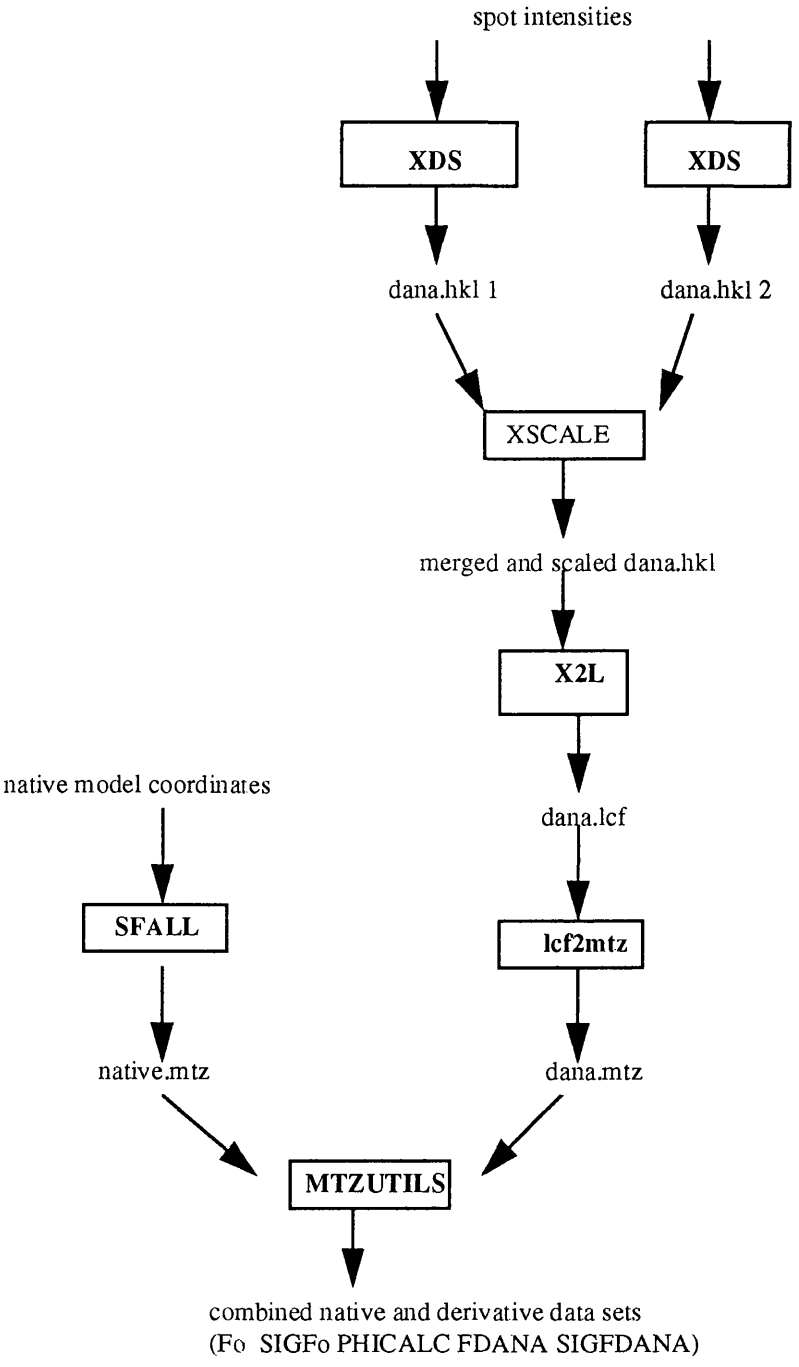
Obtaining the 2Å difference Fourier map

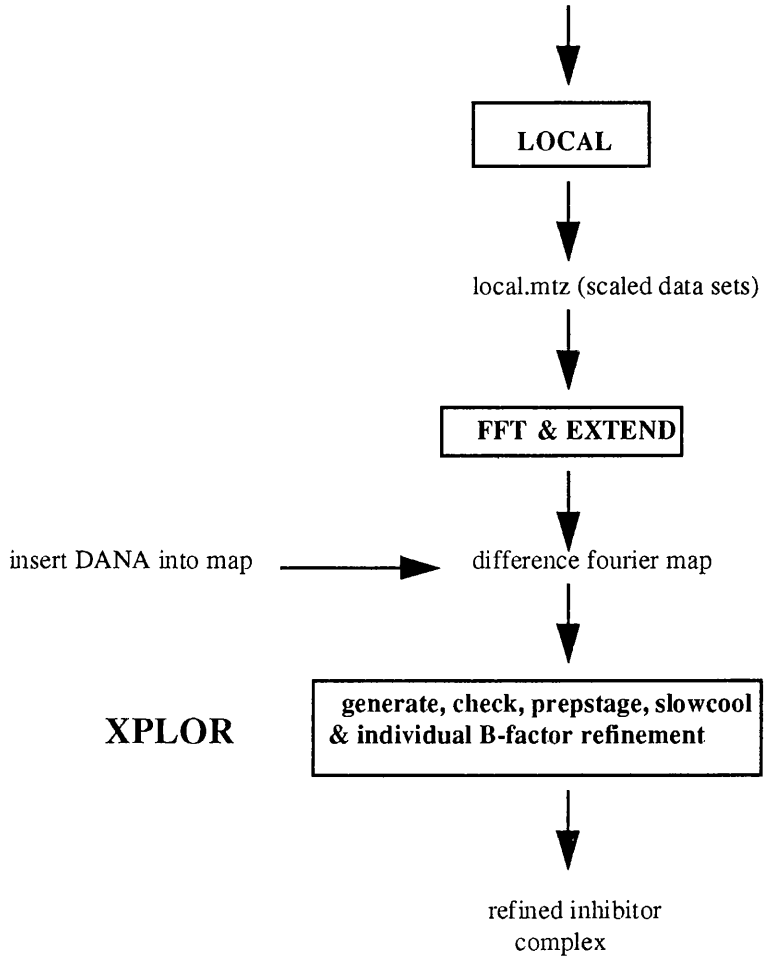
Structure factors for the final refined model of the 41kDa form of *M.viridifaciens* neuraminidase were generated to 2Å using SFALL. The derivative data set was converted to lcf format using X2L and then converted to mtz format using LCF2MTZ. Both the derivative and native mtz files were then merged using MTZUTILS and the merged reflection file scaled in LOCAL. A difference Fourier map was then generated with Fourier coefficients F_{DANA}-F_o $e^{i\alpha calc}$ using FFT. The map was extended in EXTEND to cover the region of the molecule of interest.

Building and refining the inhibitor complex structure

The difference Fourier map clearly revealed density due to scattering contributions from the inhibitor alone. Waters in, or close to the electron density were removed and the inhibitor structure was rotated and translated into the map manually, using the graphics package 'O'. Refinement of the complex was then implemented in X-PLOR. Rigid body refinement of the complex did not improve the fit of the inhibitor to the electron density and was therefore discarded. After one round of PREPSTAGE, SLOWCOOL and individual B-factor refinement with a 0σ cut-off, the R-factor was 15.9%. The refined inhibitor complex has 2 residues in generously allowed regions of the Ramachandran plot and zero residues in disallowed regions. The average main chain temperature factors were 13.45Å^2 and the average side chain temperature factors were 21.29Å^2. The average temperature factors for DANA were 12.9Å^2. A flow chart summarizing how the refined complex structure was obtained is shown in fig. 8.3.

Fig. 8.3 Obtaining the structure of the 41kDa *M.viridifaciens* - DANA complex.





B-factors in the active site

There was little change in the position of active site amino acid side chains on binding the inhibitor and the active site appears to be fairly rigid, as observed in other neuraminidases. Evidence for this comes from examining the B-factors of the active site residues. The B-factors of the active site residues in the native and inhibitor complexed structures are very similar, as shown in table 8.6. The B-factors of the terminal atoms of the active site residues in either the native or complexed structures

are significantly lower than those of a typical amino acid. The B-factor of the Nε atom of the active site residue Arg 68 in the complexed structure is 10.8Å^2, in comparison with a B-factor of 60.2Å^2 for the Nε atom of Arg 61. Similarly in the native structure, the B-factor of the OE2 atom of the active site residue Glu 386 has a value of 8.5Å^2, in comparison with 45Å^2 for the OE2 atom of Glu 48.

Table 8.6 Average B factors of the active site residues in the native and inhibitor complexed 41kDa form of *M.viridifaciens* neuraminidase.

Active site residue	**Native 41kDa enzyme (Å^2)**	**41kDa enzyme complexed with inhibitor (Å^2)**
68 Arg	11.0	10.0
92 Asp	26.8	34.0
131 Asp	12.4	12.4
276 Arg	13.4	10.6
342 Arg	11.8	13.5
370 Tyr	6.1	4.9
386 Glu	7.5	10.0
Overall average B factor (Å^2)	12.7	13.6

The Active site-interactions with the inhibitor

The similarities and differences between the interactions of the inhibitor with the active site in the 41kDa form of *M.viridifaciens* neuraminidase and in other neuraminidases are discussed in detail below. Fig. 8.4 shows the difference Fourier map, with DANA superimposed on the density. A summary of the interactions

between DANA and the active site residues in the 41kDa form of *M.viridifaciens* neuraminidase is given in table 8.7.

A) Interactions with the glycerol side chain of DANA. In *M.viridifaciens* neuraminidase, there is one strong hydrogen bond to O8 and one strong and one weak hydrogen bond to the O9 of DANA from Asp 259. Similarly, the viral enzyme has two strong hydrogen bonds to the O8 and O9 of the glycerol group of NANA and DANA, but the Asp is replaced by a Glu. Asp 637 of *V.Cholerae* neuraminidase could form interactions to the O8 and O9 and Asn 318 with O7, but since no inhibitor complex has yet been solved this can only be speculation. In contrast, *S.typhimurium* neuraminidase has only one weak hydrogen bond to the glycerol group; from Trp 128 to O9. *S.typhimurium* neuraminidase is known to have a solvent channel around the glycerol group of the inhibitor/substrate which was postulated to explain why the enzyme proceeds with inversion of configuration at the anomeric C2 of the sialic acid, since a water molecule would be able to attack underneath from the β-face of the sugar ring. However, *M.viridifaciens* neuraminidase is also thought to proceed by inversion of configuration (M. Sinnott - personal communication), yet it has stabilizing hydrogen bonds to the O8 and O9 of the glycerol group. Asp 259 must therefore be distant enough from the DANA to allow a water molecule to approach from the β-face, to initiate catalysis by inversion of configuration, whilst still maintaining hydrogen bonds to the glycerol group. There is however some conflicting evidence as to whether *M.viridifaciens* neuraminidase

does indeed proceed by inversion of configuration. This will be discussed in more detail in the subsequent section.

B) Interactions with the O4 of DANA. Interactions with the O4 of DANA in the active site of *M.viridifaciens* neuraminidase strongly resembles that in *S.typhimurium* neuraminidase and *V.cholerae* neuraminidase, but with one extra strong hydrogen bond to the O4. The O4 of DANA has strong hydrogen bonds to Asp 131, Asp 92 and Arg 87.These are replaced by Arg 245 and Asp 292 in *V.cholerae* neuraminidase and by Arg 56, Asp 62 (weakly hydrogen bonding) and Asp 100 in *S.typhimurium* neuraminidase. The presence of only a weak hydrogen bond to the O4 in the viral enzyme has enabled the rational design of potent inhbitors, which bind in this pocket.

The active site of the influenza enzyme has long been an attractive target for drug design, due to the ineffectiveness of vaccines as described in Chapter 1 and because of the invariant nature of the amino acids lining the active site. Twelve of these invariant residues (N9 subtype numbering); Arg 118, Glu 119, Asp 151, Arg 152, Trp 178, Ile 222, Glu 276, Glu 277, Arg 292, Arg 371, Asn 294 and Tyr 406 are in contact with the inhibitor DANA in the complex and a further six residues; Ser 179, Arg 224, Glu 227, His 274, Gly 348 and Glu 425 are in the second shell of active site residues, stablilzing the binding site. In the active sites of the bacterial enzymes, a conserved arginine and aspartic acid form hydrogen bonds to the O4 on the inhibitor. However these interactions are absent in the viral enzyme, resulting in a pocket close to the O4. This pocket has been filled by replacing the 4-hydroxyl on

DANA with a guanidinium group to produce a rationally designed inhibitor against the influenza enzyme (fig. 8.5). The interactions between the active site residues of influenza virus neuraminidase and bound DANA are shown in fig. 8.6.

C) Interactions with the O6 of DANA. There is one strong hydrogen bond from Tyr 370 to the O6 in the *M.viridifaciens*-inhibitor complex. *S.typhimurium* neuraminidase has one hydrogen bond to a water molecule and the *V.cholerae* enzyme may have hydrogen bonds to the O6 of DANA from Tyr 740 and Asp 250. The viral enzyme has one weak hydrogen bond to the O6 from the OE1 of Glu 279 and one strong hydrogen bond from Tyr 406.

D) Interactions with the Carboxylate group of DANA. Interactions between the carboxylate group of the inhibitor and the active site are common to all neuraminidase structures studied to date. Three arginines, Arg 68, Arg 276 and Arg 342, form hydrogen bonds in a plane with the carboxylate group of the inhibitor. Arg 68 is stabilized by Glu 386 through a hydrogen bond between the Nε of the Arginine and Oδ1 of the Glutamic acid. Tyr 370 and Glu 260 hydrogen bond with each other and sit beneath and close to the C1-C2 bond of the inhibitor.

E) Interactions with the O7 of DANA. The O7 of DANA makes two strong hydrogen bonds to two water molecules and a weak hydrogen bond to one further water molecule. As with the viral and *S.typhimurium* enzymes there are no

interactions with protein atoms. However the *V.Cholerae* enzyme may make hydrogen bonds to Asp 250 and Asn 318.

F) Interactions with the N5 of DANA. The N5 atom of DANA makes one strong hydrogen bond to the OD1 atom of Asp 131. Similarly the *S.typhimurium* enzyme makes a hydrogen bond from the OD2 atom of Asp 100 to N5 and the *V.Cholerae* enzyme may form a hydrogen bond between the N5 and Asp 250. The viral enzyme makes no interactions with the N5 atom of DANA.

G) Interactions with the O10 of DANA. The O10 of DANA makes one strong hydrogen bond to a water molecule. In contrast the viral enzyme makes hydrogen bonds to Asp 152 and Arg 153. In the *S.typhimurium* enzyme, the O10 interacts with one water molecule. The *V.Cholerae* enzyme may make hydrogen bonds to Asp 250 and Asn 318.

Table 8.7 Summary of the interactions between the bound inhibitor DANA and active site residues in the 41kDa form of *M.viridifaciens* neuraminidase. Equivalent residues in the Influenza, *V.Cholerae* and *S.typhimurium* enzymes are given and differing equivalent residue types are given.

DANA atom	mv[1] atom	distance (Å)	Equivalent residues in: st[2]	vc[3]	influenza[4]
O1A	Tyr 370 OH	3.35	342 OH	740 OH	406 OH
	Arg 68 NH1	2.76	37 NH2	224 NH1	118 NH2
	Arg 68 NH2	2.97	37 NH1	224 NH2	118 NH1
	Arg 342 NH1	2.85	309 NH1	712 NH1	371 NH2
O1B	Tyr 370 OH	3.37	342 OH	740 OH	406 OH
	Arg 276 NH1	3.59	246 NH2	635 NH2	292 NH1
	Arg 276 NH2	3.21	246 NH1	635 NH1	292 NH2
	Arg 342 NH1	3.50	309 NH1	712 NH1	371 NH2
	Arg 342 NH2	2.93	309 NH2	712 NH2	371 NH1
O4	Asp131 OD2	2.75	100 OD1	250 OD2	Glu 120 OE1
	Asp 92 OD1	2.60	62 OD2	Arg 245 NH1	
	Arg 87 NH2	3.11	56 NH2	245 NH2	
	water 581 O1	3.30	-	-	
O6	Tyr 370 OH	3.09	-	740 OH	Tyr 406 OH
	water 773 O1	3.40	690 O1	Asp 250 OD1	Glu 279 OE1
O7	water 773 O1	2.90	689-691	Asn 318 ND2	-
	water 772 O1	3.57	O1	Asp 250 OD1	
	water 774 O1	3.01			
O8	Asp 259 OD1	2.69	water 694	Asp 637 OD2	Glu 279 OE1
			& 688 O1	Glu 619 OE2	Glu 278 OE2
O9	Asp 259 OD2	2.90	Trp 128 Nε1	-	Glu 278 OE1
O10	water 582 O1	2.69	689	Asp 250 OD1	Asp 152 OD1
			O1	Asn 318 OD1	Arg 153 NH1
					Arg 153 NE
N5	Asp 131 OD1	2.93	100 OD2	250 OD1	-

[1] *M.viridifaciens* neuraminidase

[2] *S.typhimurium* neuraminidase

[3] *V.cholerae* neuraminidase. N.B There is no crystal structure for the neuraminidase from *V.cholerae* complexed with inhibitor. The interactions given above for *V.cholerae* with DANA are the results of modelling studies only.

[4] Influenza virus neuraminidase subtype N9.

Fig. 8.4 Stereo view of the difference Fourier map at 2Å, with coefficients F_{DANA}-$F_o e^{i\alpha calc}$ and the refined coordinates of DANA superimposed on the density. The map is contoured at 3.5σ.

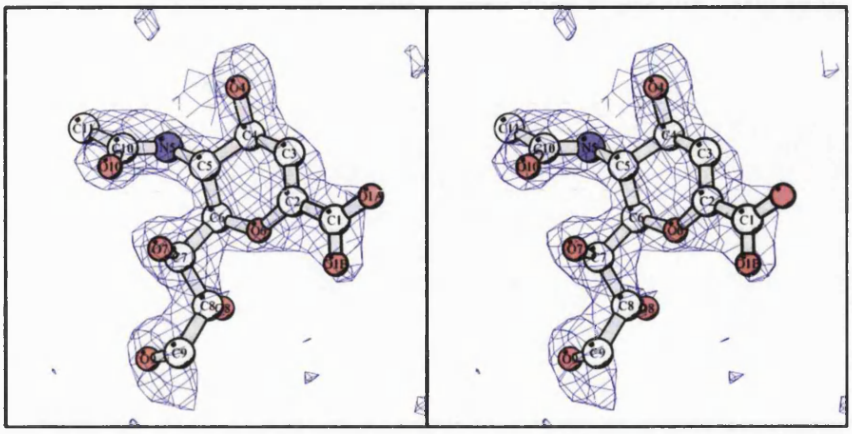


Fig. 8.5 4-guanidino-Neu5Ac2en. A potent rationally designed inhibitor against influenza neuraminidase.

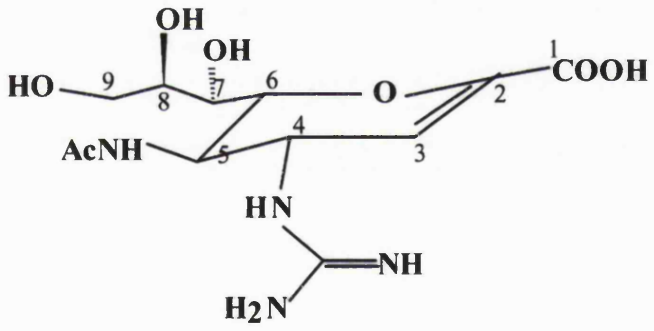
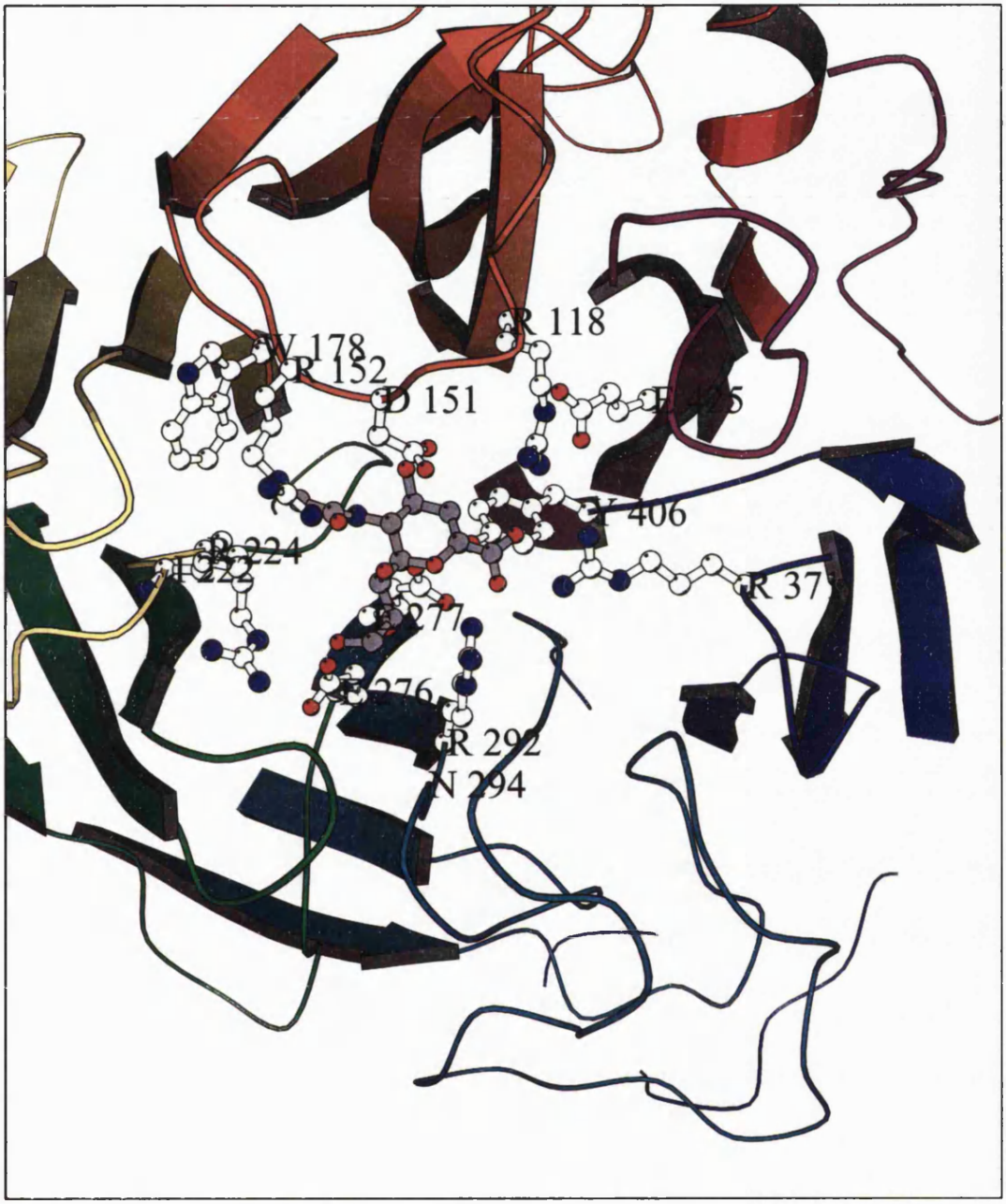


Fig. 8.6 Interactions between the active site residues of influenza virus neuraminidase (subtype N9 numbering) and the inhibitor DANA. Some active site residues have been omitted for clarity.



The differences between the *M.viridifaciens* and the other bacterial and viral neuraminidase active sites seem to indicate that the sialic acid in *M.viridifaciens* neuraminidase is bound more tightly than in any of the other neuraminidases. It was previously suggested that a possible explanation for a much higher turnover in salmonella as compared with the viral enzyme is the stabilization of the glycerol chain by hydrogen bonds in the viral enzyme, product release being the limiting factor in catalysis (Crennell *et al.*, 1993). However the bound DANA in *M.viridifaciens* neuraminidase is located deeper in the active site than in the salmonella enzyme, and has more stabilizing hydrogen bonds, yet the *M.viridifaciens* neuraminidase has a high activity; being at least 200 units/mg protein. It is therefore unlikely that product release is the rate limiting factor in catalysis by *M.viridifaciens* neuraminidase.

Substrate specificity and enzyme mechanism

In common with many of the bacterial neuraminidases *M.viridifaciens* neuraminidase can cleave α2-3, α2-6, and α2-8 linked sialic acids. In contrast to the other bacterial neuraminidases *S.typhimurium* neuraminidase has a 260-fold kinetic preference for α2-3 linkages over α2-6 linkages similar to the influenza enzyme. This variation in substrate specificity appears to be due to steric hindrance contributions alone arising from the variability of the lengths of the loops around the active site. Table 8.8 shows the substrate specificity for those viral and non-viral neuraminidases whose structures have been determined.

Table 8.8 The variable substrate specificity of the neuraminidases. Values are given as relative cleavage rates (%). Only those neuraminidases whose structures have been determined are included.

	α(2-3)	α(2-6)	α(2-8)	references
M.viridifaciens	100	253	125	Sakurada et al., 1992
V. cholerae	100	53	30	Corfield et al., 1993
S.typhimurium	100	<1	<1	Hoyer et al., 1991
Influenza virus A	100	<1	3	Corfield et al., 1993, Corfield et al., 1982

Substrates used were: α(2-3), Neu5Acα(2-3)Galβ(1-4)Glc; α(2-6), Neu5Acα(2-6)Galβ(1-4)Glc; α(2-8), colominic acid.

The mechanism of catalysis of *M.viridifaciens* neuraminidase has been investigated using optical rotation and kinetic isotope effects. Like *S.typhimurium* neuraminidase, *M.viridifaciens* neuraminidase also proceeds by inversion of configuration at the anomeric C2 of the substrate (Guo & Sinnott, 1993 and M. Sinnott - personal communication). In contrast it has been shown that the neuraminidases from *V.cholerae* (Guo & Sinnott, 1993) and influenza (Chong *et al.*, 1992) proceed by overall retention of configuration. The situation is unclear however, since a recent nmr study of *S.typhimurium* neuraminidase showed that the enzyme catalysed the cleavage of sialic acid with overall retention of configuration (Wilson *et al.*, 1995). In addition other nmr studies on the neuraminidase from *M.viridifaciens*, *S.typhimurium* and the small *C.perfringens* isoenzyme show a

mixture of retention and inversion of configuration (M.Sinnott - personal communication), whereas the *V.cholerae* neuraminidase is definitely retentive.

The mechanism of catalysis by influenza virus neuraminidase has been studied in most detail. Kinetic isotope methods (Chong *et al.*, 1992), nmr (Chong *et al.*, 1992; Janakiraman *et al.*, 1994), and a molecular dynamics simulation of the enzyme-complex (Taylor & von Itzstein, 1994) have been used in order to further an understanding of the enzyme mechanism. A scheme has been proposed involving an endocyclic sialosyl cation intermediate (Chong *et al.*, 1992). Fig. 8.7 summarizes a possible mechanism for hydrolysis of the substrate by the neuraminidase from influenza virus. The binding of the sialic acid to the neuraminidase involves considerable distortion of the pyranose ring but the extensive hydrogen bonding network to the substrate compensates for the free energy lost. In solution, the sialic acid is in the chair conformation, but upon binding the pyranose ring adopts a pseudoboat conformation. The planar nature of the sugar around the carboxylate carbon and the ring atoms O6, C2 and C3 activates the sialic acid for hydrolysis. At the optimum pH for catalysis by influenza virus neuraminidase, the hydrolysis of the substrate proceeds with substantial positive charge developing at the reaction centre in the transition state for the formation of the sialosyl-enzyme intermediate. The binding site of the substrate involves a protonated Arginine. The positive charge which develops in the sialosyl cation transition state could be stabilized by the neighbouring negative charge of the carboxylates. The catalysis is probably solvent mediated and may involve water hydrogen-bonded to an Aspartic acid in an enzyme mediated proton transfer in a S_N1-type mechanism. The product of the reaction,

sialic acid, is initially released as the α-anomer which converts to the more favourable β-anomer in solution by mutarotation (fig. 8.8).

Fig. 8.7 Proposed mechanism for catalysis by neuraminidase from influenza virus (Chong *et al.*, 1992).

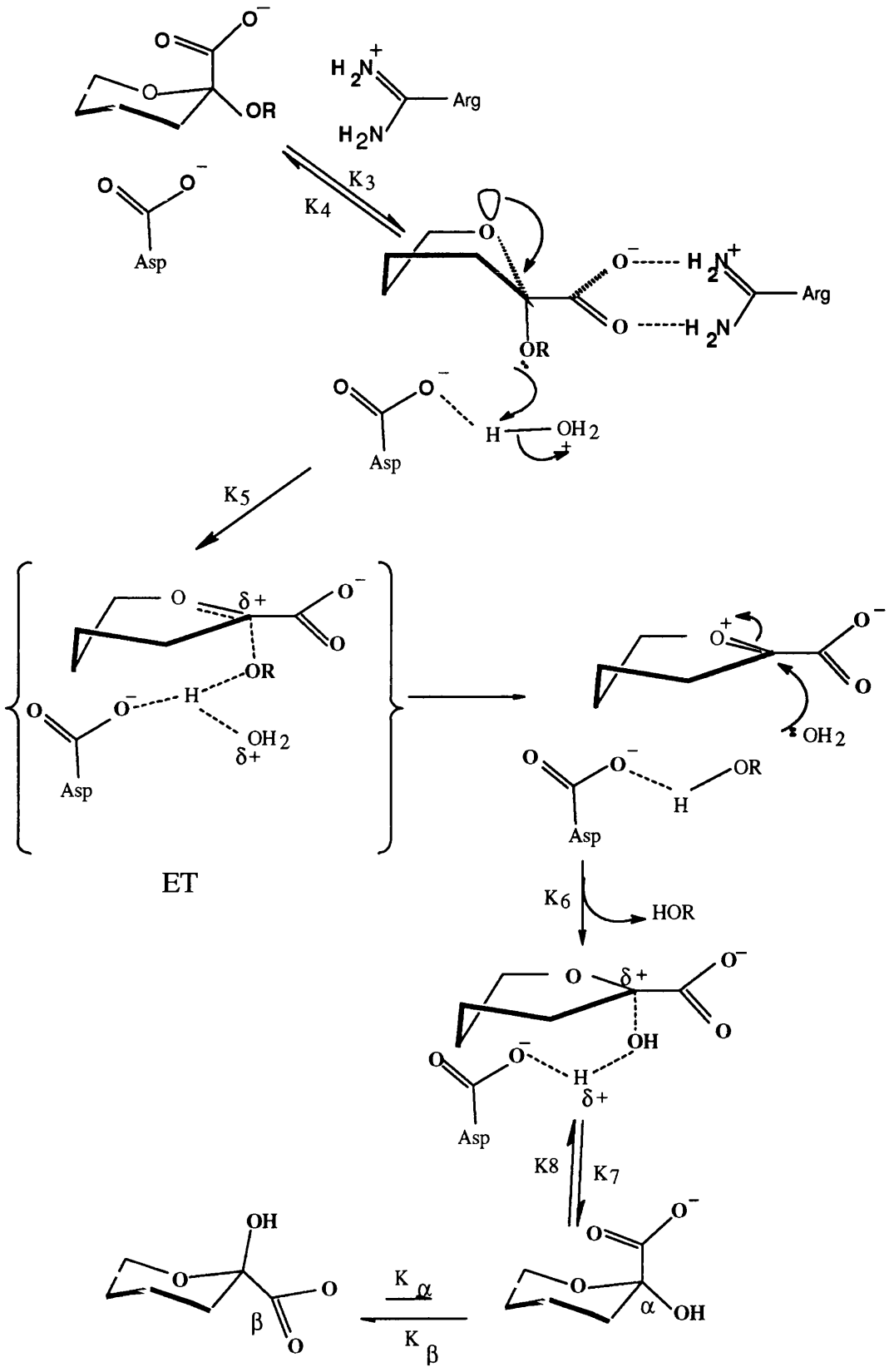
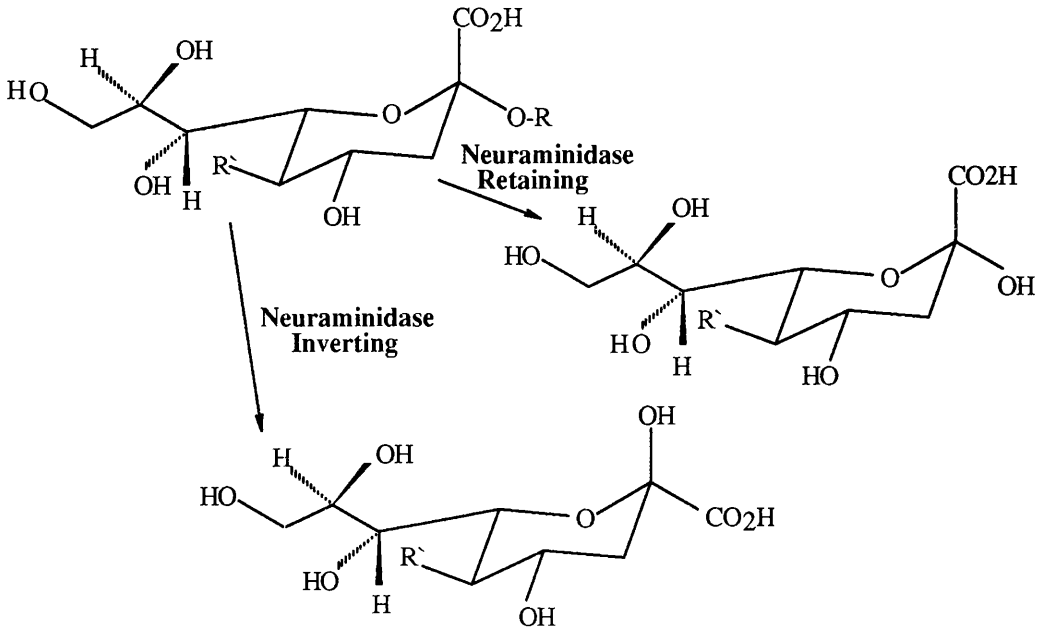


Fig. 8.8 Mutarotation of N-acetyl neuraminic acid.



Chapter 9

THE 68kDa *M.viridifaciens* NEURAMINIDASE STRUCTURE

Protein architecture

The 68kDa form of *M.viridifaciens* neuraminidase (fig. 9.1) possesses three domains (fig. 9.2). The 41kDa β-propeller domain, a domain with the immunoglobulin fold and a domain with high homology to the galactose binding domain of galactose oxidase from *Dactylium dendroides*. The 41kDa domain of 68kDa *M.viridifaciens* neuraminidase shows very little difference to the 41kDa enzyme in isolation (Chapter 8) and the two 41kDa domains have a rmsd of 0.4Å over 358 Cα atoms when superimposed using the program MNYFIT in the COMPOSER suite of software. The only significant structural change between the 41kDa domain of the 68kDa form of *M.viridifaciens* neuraminidase and the isolated 41kDa form is in residues 402-404 which move to connect to the linker domain.

Fig. 9.1 The 68kDa form of *M.viridifaciens* neuraminidase. Domain 1 is shown in mustard, domain 2 in burgundy and domain 3 in olive.

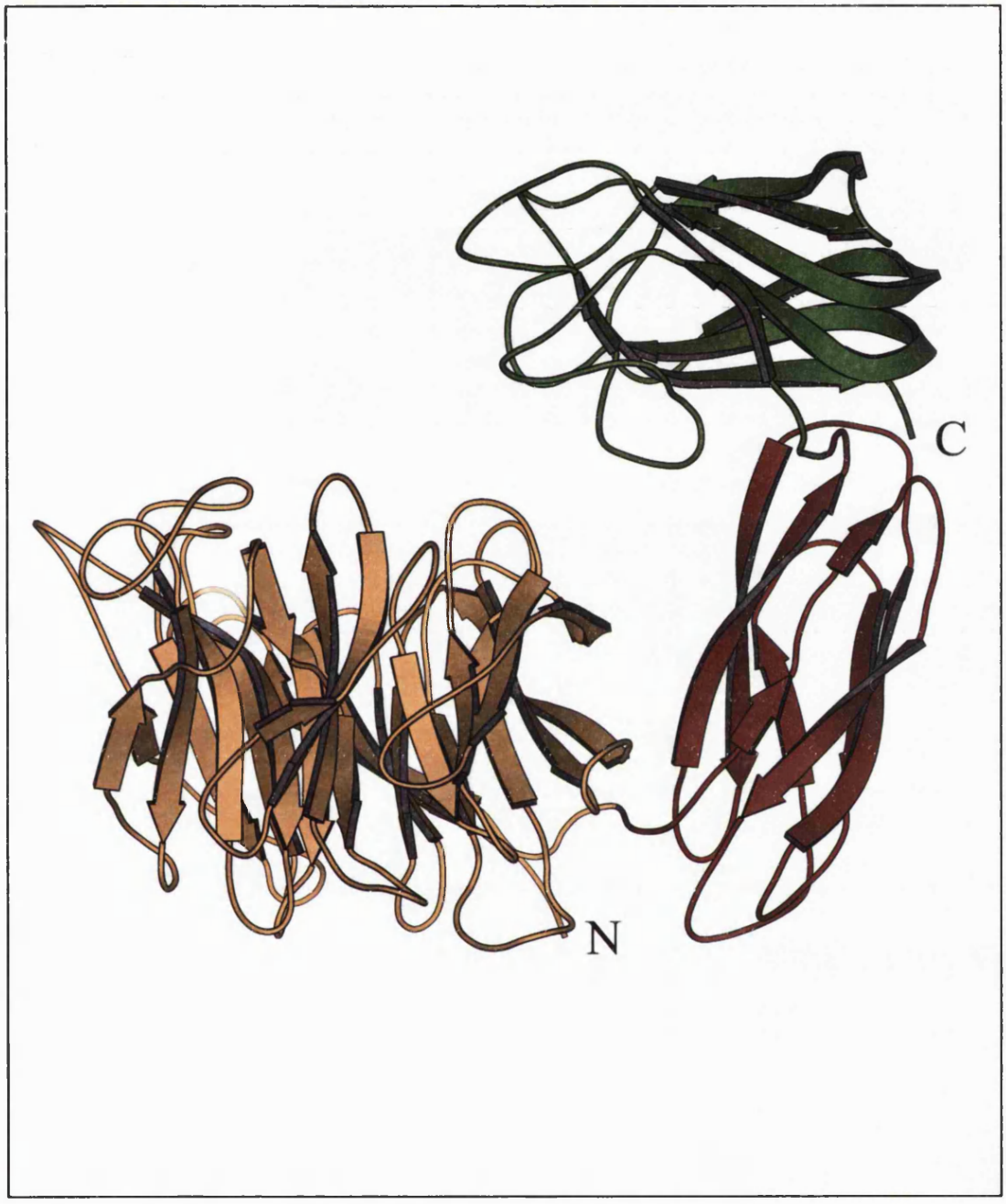


Fig. 9.2 The three domains of 68kDa *M.viridifaciens* neuraminidase. Domain 1 (a) is the 41kDa β-propeller domain, domain 2 (b) is the linker domain with the immunoglobulin fold and domain 3 (c) is the galactose binding domain. The view is orthogonal to fig. 9.1.

The linker domain

The second domain leaves the 41kDa domain at residue 405 and enters the galactose binding domain at residue 506. This domain, termed the linker domain, consists of a β-sandwich of one β-sheet with 4 antiparallel strands over one β-sheet with 3 antiparallel strands (Fig. 9.2 (b)) and has a well-packed hydrophobic interior. Following a search using the program PROTEP (P. Artymuik personal comm.) the fold of the linker domain was found to most resemble that of the immuoglobulin family, particularly the immunoglobulin variable domain. Both the variable and constant immunoglobulin domains have been described as a Greek key β-barrel, subclass "simple" (Richardson, 1981). The immunoglobulin structural domains have between 7 and 9 antiparallel β-strands which form a barrel. The barrel is divided into two distinct β-pleated sheets since the hydrogen bonds do not go around the barrel and so the barrel appears as a β-sandwich. However in contrast to the majority of immunoglobulin folds, the linker domain forms an almost cylindrical 7-stranded β-barrel structure, which is stablilised by the first strand (strand a in immunoglobulin nomenclature (Bork *et al.*, 1994)) forming hydrogen bonds with strands b and g. Strand a is able to hydrogen bond both strands b and g by having a proline halfway along (Pro 411), which changes the strand direction from the b strand to the g strand on the opposite sheet. The linker domain does not contain any disulphide bonds and its sequence does not possess the consensus shown by immunoglobulin folds. The fold of the linker domain is therefore another variation on that of the immunoglobulins. 2-D topology diagrams illustrating the observed hydrogen-bonding pattern and strand topology in the immunoglobulins are shown in fig. 9.3. Using the program PROTEP,

structural homologies to the linker domain were also found with the C-terminal domain of CD4 and the N-terminal domain of the chaperone protein PapD (fig. 9.4). The latter homology is interesting in that the strand a of PapD also "dog-legs" to form hydrogen bonding to both strands b and g, leading to a more cylindrical barrel as found in the linker domain (Holmgren & Branden, 1989).

Domains with the immunoglobulin-like fold have been reported in a number of non-immunoglobulin structures, including matrix proteins e.g. tenascin (Leahy *et al.*, 1992b), cell surface receptors e.g. CD2 (Driscoll *et al.*, 1991; Jones *et al.*, 1992), CD8 (Leahy *et al.*, 1992a) and interestingly galactose oxidase from *Dactylium dendroides.* With the large number of structures falling into the class of the immunoglobulin fold and with the huge number of proteins with sequences related to that of proteins having the immunoglobulin fold, the immunoglobulin-like domain is probably the most widespread protein module, at least in animals (Doolittle & Bork, 1993).

Fig. 9.3 2-D topology diagrams illustrating the observed hydrogen-bonding patterns in the immunoglobulins. Immunoglobulin constant domains have 7 strands in the topology shown as C-type. The variable domain, V-type, has an additional hairpin c`-c`` between strands c and d, with strand a having two alternative locations. Strand a can either be part of the back sheet (antiparallel pairing with strand b), or part of the front sheet (parallel pairing with strand g). A switched, S-type, fold can also occur. Here the 4th strand has switched sheets (d to c`). An 8 stranded hybrid, h-type, between types C and S is also possible. This has both strands c` (front sheet) and its direct continuation strand d (back sheet), so that both sheets have 4 strands. The linker domain of 68kDa *M.viridifaciens* neuraminidase has a fold most like that of the V-type immunoglobulins.

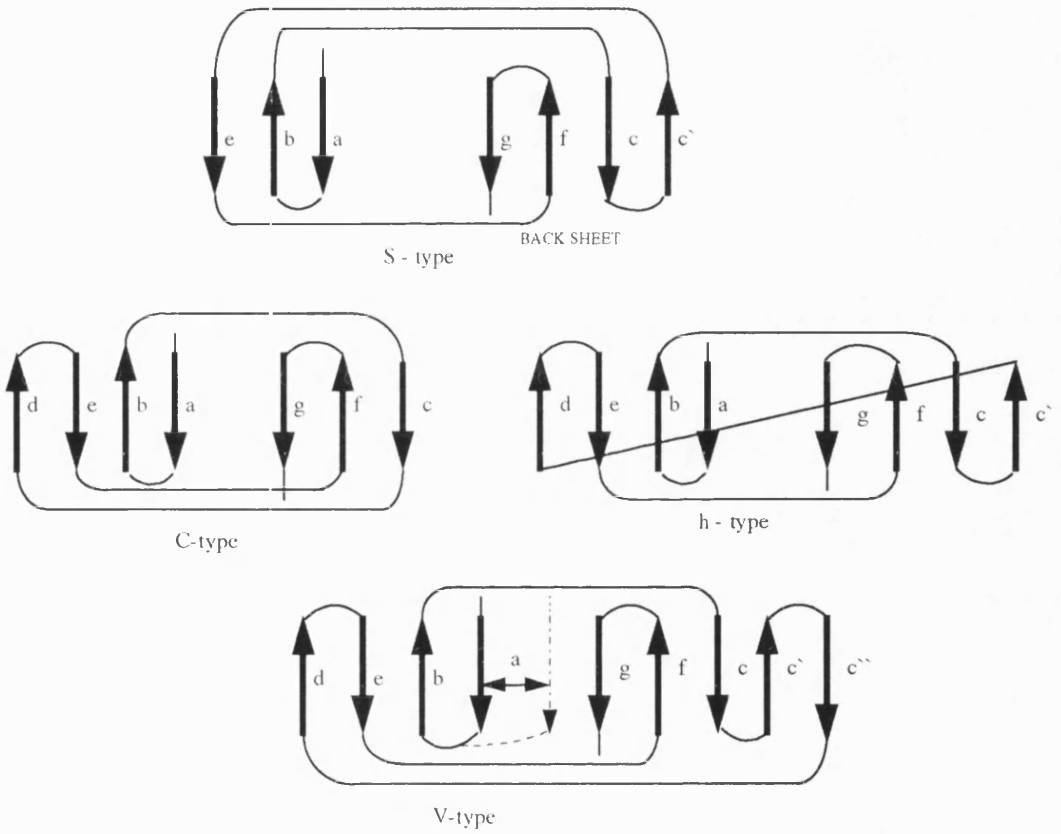
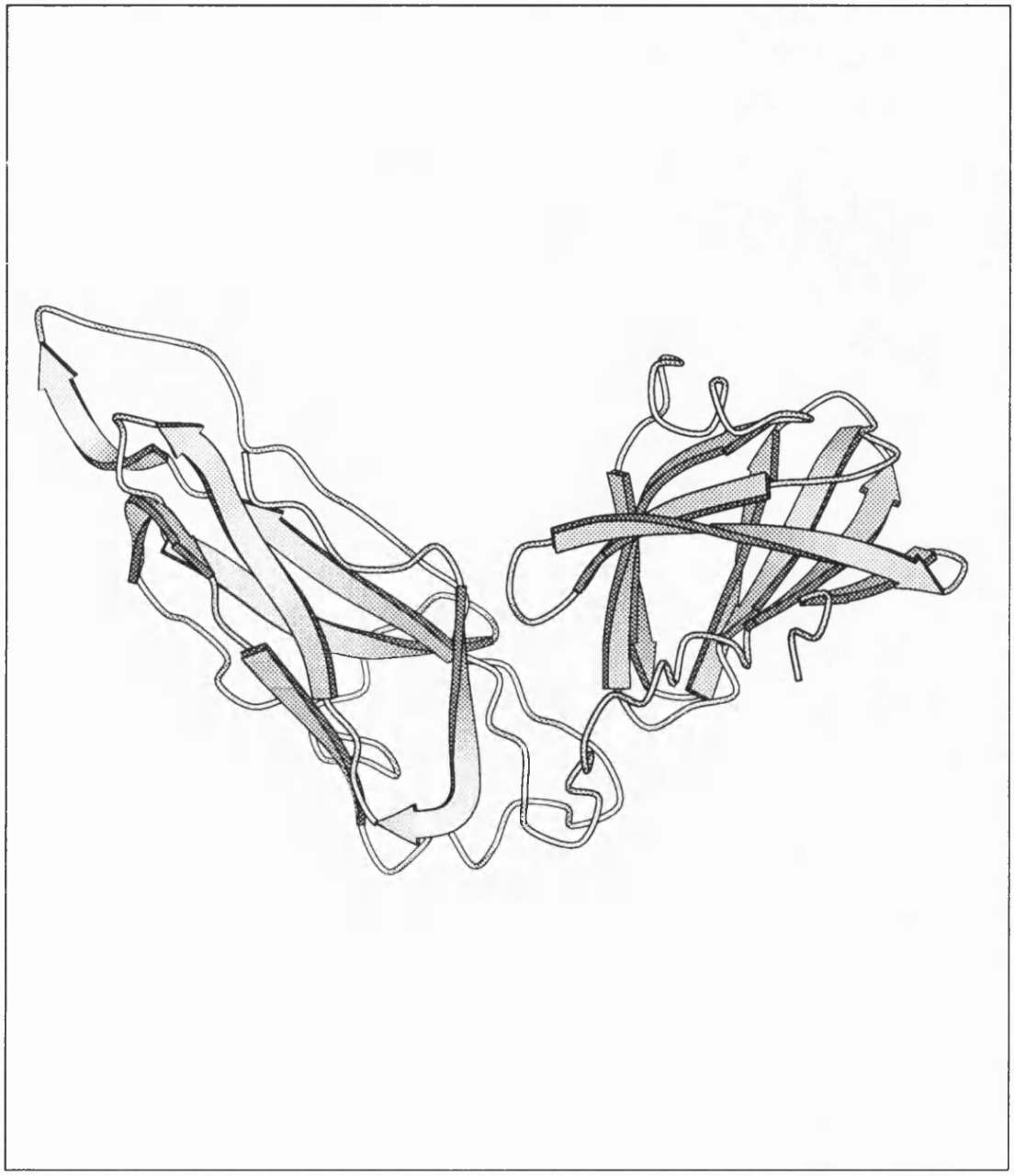


Fig. 9.4 MOLSCRIPT drawing of the chaperone protein PapD. The two immunoglobulin domains are each composed of two β-sheets, with strands a b e forming one β-sheet and strands c f g forming the second β-sheet.



The Galactose binding domain

The polypeptide backbone leaves the linker domain at residue 506 and enters the final domain which extends to residue 647. This latter domain has been termed the

galactose binding domain due to its high structural and sequence homology with domain 1 of galactose oxidase from *Dactylium dendroides*. As summarized in Chapter 7, galactose oxidase consists of three domains. Domain 1 has the jelly-roll motif and consists of a β-sandwich, with a five-stranded antiparallel β-sheet facing a three-stranded antiparallel β-sheet. Domain 2 contains 28 β-strands and one short α-helix. The residues implicated in catalysis are located in this domain. This domain has the β-propeller fold, with pseudo 7-fold symmetry as compared with the 6-fold symmetry found in the neuraminidases. The C-terminal domain 3 of galactose oxidase consists of a bundle of seven β-strands which make a complicated hydrogen bonding network surrounding a hydrophobic core.

The galactose binding domain of 68kDa *M.viridifaciens* neuraminidase has the same fold as domain 1 of galactose oxidase, consisting of a β-sandwich fold where a five-stranded antiparallel β-sheet faces another three stranded antiparallel β-sheet (fig. 9.2 (c)). The sequence homology between the galactose binding domains from the two enzymes has been discussed in Chapter 7 and upon superposition using MNYFIT in COMPOSER, the two domains gave a rmsd of only 1.13Å over a superposition of 136 Cα atoms.

Domain 1 of galactose oxidase is known to have a binding site for a small cation. This cation was assigned a sodium ion in galactose oxidase since it was the abundant cation in solution (Ito *et al.*, 1994). However the binding site is not very specific for the cation since it was observed that other cations could replace the sodium ion (Ito *et al.*, 1994). The presence of a cation binding site in the galactose binding domain of the 68kDa form of *M.viridifaciens* neuraminidase was also established (fig. 9.6). The binding site is located in a very similar position to that in the galactose

oxidase, and the ligands are the same. The cation maintains the conformation of the loop (residues 528-536) and the coordination to the ligand is a pentagonal bipyramidal configuration, as in galactose oxidase. A comparison of the ligands for the cation is given in table 9.1. Fig. 9.5 shows a sequence alignment of galactose oxidase domain 1 with domain 3 of the 68kDa form of *M.viridifaciens* neuraminidase, highlighting those residues involved in cation and galactose binding in galactose oxidase which are conserved in the neuraminidase. The alignment shown here was produced using the BESTFIT program in the GCG suite of software and as such differs from the alignment shown in Chapter 7, Fig. 7.3, which was produced using the program FASTA. However the two different programs did not alter the way in which the residues involved in cation and galactose binding aligned.

Fig. 9.5 Sequence alignment of domain 1 of galactose oxidase with domain 3 of the neuraminidase. Residues involved in cation and galactose binding in galactose oxidase are boxed and shaded respectively. These residues are mostly conserved in the neuraminidase.

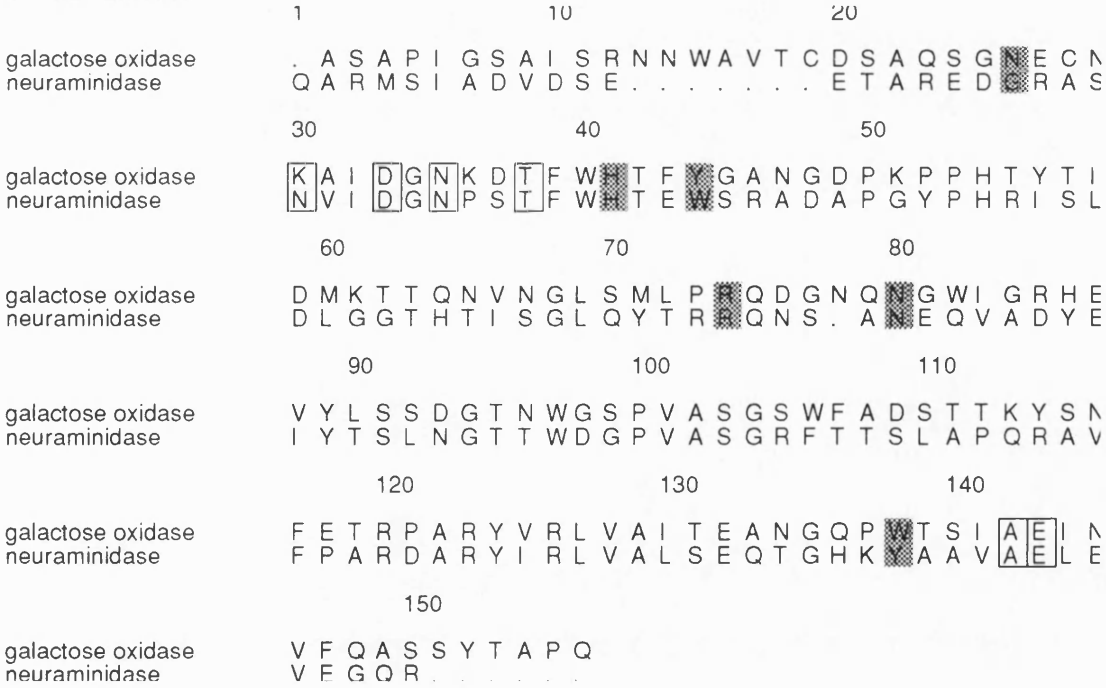


Table 9.1 (a) Ligands for the cation in galactose oxidase.

First shell of ligands (2.0 - 2.4Å)

residue	atom	ion - ligand distance (Å)
Lys 29	main chain oxygen	2.14
Thr 37	main chain oxygen	2.01
Asn 34	main chain oxygen	2.22
Ala 141	main chain oxygen	2.36
Asp 32	OD1	2.28

Second shell of ligands (~2.6Å)

residue	atom	ion - ligand distance (Å)
Thr 37	OG1	2.61
Glu 142	OE2	2.62

(b) Ligands for the cation in the galactose binding domain of the neuraminidase.

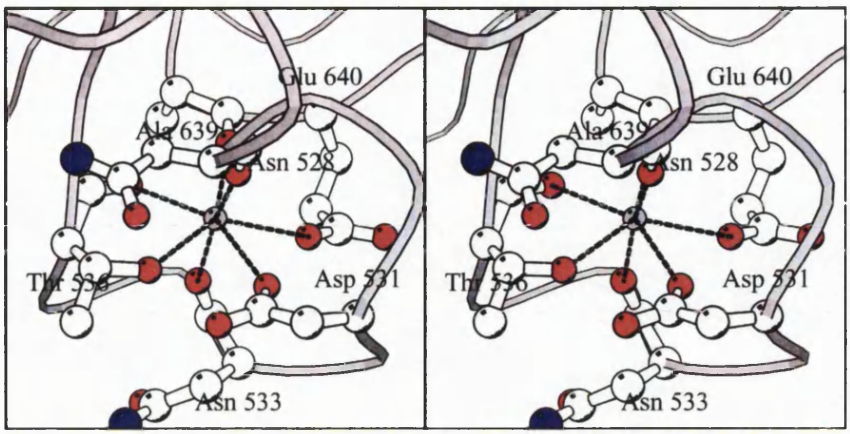
First shell of ligands (2.0 - ~2.4Å)

residue	atom	ion - ligand distance (Å)
Asn 528	main chain oxygen	2.12
Thr 536	main chain oxygen	2.48
Asn 533	main chain oxygen	2.36
Ala 639	main chain oxygen	2.32

Second shell of ligands (~2.6Å)

residue	atom	ion - ligand distance (Å)
Asp 531	OD1	2.67
Glu 640	OE1	2.59
Thr 536	OG1	2.55

Fig. 9.6 Stereo view of the cation binding site in the galactose binding domain of 68kDa *M.viridifaciens* neuraminidase.



Domain 1 of galactose oxidase was known to bind galactose with low affinity (Ito *et al.*, 1994). However this galactose binding site is not likely to be the active site, being some 40Å from the copper site and is therefore a secondary site. Hence in order to see whether the galactose binding domain of the 68kDa form of *M.viridifaciens* neuraminidase also possessed an affinity for galactose, a crystal of the native protein was soaked in a solution of 50mM galactose for 4 hours and a data set collected.

Collecting data from the 68kDa crystal soaked in galactose

Data was collected using the same native "magic" 68kDa crystal which had previously yielded 2.5Å data (Chapter 3), but soaked in 50mM galactose for 4 hours. Data was collected on station X31 at the DESY synchrotron, Hamburg, on station X31 at a wavelength of 0.993Å. Data was collected in 1° frames with exposure times

between 170 and 300 seconds. Diffraction extended to 2.5Å, and was 62.1% complete (>1σ). Data collection statistics are shown in table 9.2.

Obtaining the 2.5Å difference Fourier map and refining the enzyme-galactose complex

The 2.5Å difference map for the galactose was generated in the same manner as for the inhibitor 2-deoxy-2,3-dehydro-N-acetylneuraminic acid (DANA) with the 41kDa *M.viridifaciens* neuraminidase (Chapter 8). With the difference map contoured at 3σ the presence of difference density due to contributions from the galactose alone clearly revealed the galactose binding site. The galactose structure was manually rotated and translated onto the difference map using the graphics package 'O' in order to find the best fit of the molecule to the density. Refinement of the complex was then implemented in X-PLOR. Rigid body refinement of the complex reduced the R-factor from 31.2% to 28.9%. A single round of least squares refinement in PREPSTAGE with a 0σ cut-off gave a final R-factor of 22.8%. Fig. 9.7 shows the refined position of the galactose superimposed on the difference electron density.

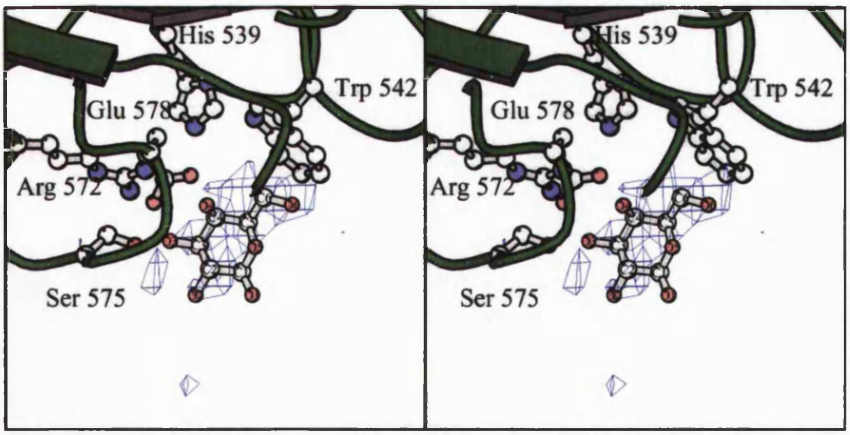
Table 9.2 Data collection statistics for the galactose-enzyme complex.

resolution (Å)	no. of unique reflections	completeness (>1σ) %	completeness (>3σ) %	R_I (%)[a]
6.76 - 5.38	788	60.6	60.4	3.7
4.27 - 3.97	829	64.0	63.3	4.4
3.55 - 3.39	831	64.6	63.3	5.6
3.15 - 3.05	835	65.8	64.2	6.8
2.89 - 2.82	877	66.4	61.9	8.3
2.69 - 2.64	874	65.9	60.4	11.3
2.54 - 2.5	366	65.9	9.5	13.5
∞ - 2.5	16304	62.1	58.6	5.3

cell dimensions	a	b	c	α	β	γ
	50.87	116.99	60.00	90.00	95.60	90.00

$$^{a}R_I = \frac{\sum_h \sum_i |(I_h - I_{hi})|}{\sum_h \sum_i I_{hi}}$$

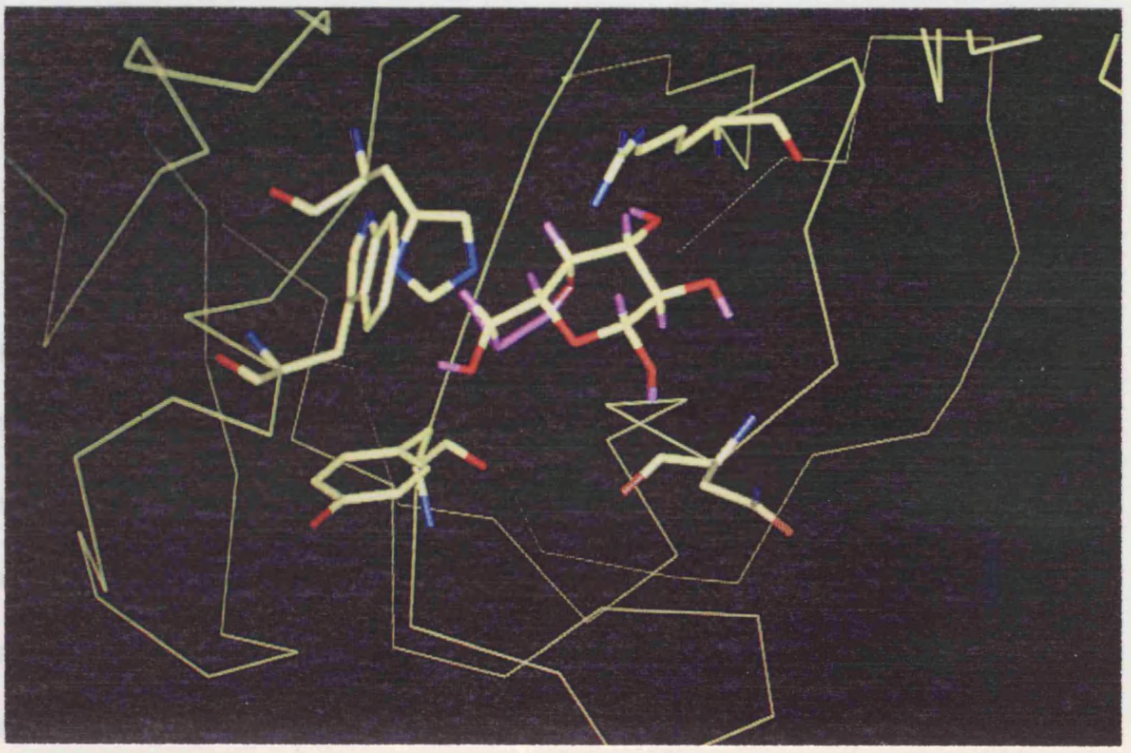
Fig. 9.7 Stereo view of the difference Fourier map at 2.5Å, with coefficients $F_{galactose}$-$F_o e^{i\alpha calc}$ and with the refined coordinates of the galactose superimposed on the density. The map is contoured at 3σ.



Analysis of the sugar binding site

The wall of the galactose binding pocket in galactose oxidase consists of a hydrophilic half and a hydrophobic half. Tyr 43 and Trp 137 in the hydrophobic half interact with the carbon backbone of galactose. Asn 25, His 40, Arg 73 and Asn 79 in the hydrophilic half form several hydrogen bonds to the O3 and O4 of the galactose. In the neuraminidase, interactions between the galactose and binding site residues are subtly different (fig. 9.8). In the hydrophobic half, Tyr 43 is replaced by Trp 542, which makes hydrogen bonds from the carbonyl oxygen to the O6 of galactose. Trp 137 is replaced by Tyr 635, which is too distant to form hydrogen bonds to the galactose. In the hydrophilic half Asn 25 is replaced by Arg 525, quite distant from the galactose, His 40 is replaced by His 539 making no hydrogen bonds, Arg 73 is replaced by Arg 572 which makes hydrogen bonds to the O3 and O4 of galactose and Asn 79 is replaced by Asn 577, too distant to make hydrogen bonds.

Fig. 9.8 Interactions between bound galactose and ligands in the 68kDa form of *M.viridifaciens* neuraminidase. Although the residues within the binding pocket are conserved with galactose oxidase, the ligands differ.



Mobile modules

It has been recognized for some time that many proteins consist of a small set of modular elements. A domain (module) with very similar fold is often found in many different proteins. By creating new proteins with the components of other ones, the potential for protein diversity could increase. Rossmann first realized this when he solved the structure of lactate dehydrogenase (Rossmann *et al.*, 1974). He discovered that the co-factor binding domain had counterparts in other dehydrogenases and these became known as having the Rossmann fold. Rossmann also noticed that these homologous domains did not necessarily occur in the same relative parts of the molecules, suggesting gene shuffling events. As a result Rossmann proposed the theory of mobile modules; in that proteins were composed of domains (or modules) that had

evolved early in life's history. By moving as discrete units from one protein to another during evolution, these mobile modules increase protein diversity. The number of domains being recognized as mobile modules is increasing and well documented examples include the immunoglobulin family, fibronectin types 1, 2 and 3 and the lectin family. The linker and galactose binding domains of 68kDa *M.viridifaciens* neuraminidase are examples of such mobile modules.

The similarity of the sequences from the non-catalytic domains of 68kDa *M.viridifaciens* neuraminidase and the amino terminal domain of galactose oxidase has already been reported (Bork & Doolittle, 1994) (Chapter 7). However there is no sequence resemblance between the catalytic domains of the two enzymes, in spite of the fact that both are now known to have similar folds; domain 2 of galactose oxidase has a 7-sheet β-propeller. This suggests that the non-catalytic domains of *M.viridifaciens* neuraminidase were shuffled into their present location after the divergence of the catalytic domain. Following elucidation of the crystal structure of 68kDa *M.viridifaciens* neuraminidase, the non catalytic domains were revealed to consist of a domain with the immunoglobulin fold and a domain within which the sequence similarity to galactose oxidase lay.

The immunoglobulin fold strongly resembles that of fibronectin type III domains, another class of mobile modules. This domain is often found in animal proteins, but is also found in bacterial proteins. The enzyme occurs as a distinct seven-stranded β-sandwich. The fibronectin type III domain has been well studied (Little *et al.*, 1994, Doolittle & Bork, 1993). It appears from phylogenetic studies that bacteria have acquired Fn3 units by horizontal transfer between themselves (Little *et al.*, 1994). This has been taken further with the proposal that the initial acquisition of the Fn3 unit

in bacteria was from a eukaryotic source, via transformation. A scenario may be envisaged whereby a soil bacterium is transformed by the incidental acquisition of a piece of DNA from some decaying organism. A segment of that DNA may encode a domain which confers some natural advantage on the organism and so after the initial incorporation of the DNA, in time the lineage may become the major strain or even take over the genus. Support in favour of the general process of transfer of genetic material between prokaryotes and eukaryotes has been given by Smith *et al.* (1992) and in the case of the transfer of the Fn3 domain a body of evidence exists (Little *et al.*, 1994). Interestingly, this evidence includes the fact that the distribution of the Fn3 domain is at present limited to animals and soil bacteria. Also it may be that soil bacteria are especially susceptible to horizontal interkingdom transfer. Enteric bacteria are clonal in their evolution whereas soil bacteria have heterogeneous genomes reflecting greater amounts of recombination. The organism *M.viridifaciens* is known to be a soil living organism and as such would have had the potential in the past to acquire such a mobile module by horizontal gene transfer from a eukaryote.

It is therefore possible that *M.viridifaciens* has acquired immunoglobulin-like and galactose binding modules from eukaryotes such as fungi or animals. Galactose oxidase and *M.viridifaciens* neuraminidase may share a common ancestor in a primitive 4-stranded β-sheet, which through gene duplication and functional divergence has produced two domains with different catalytic functions.

Chapter 10

CONCLUSIONS AND FURTHER WORK

The elucidation of the structure of the 41kDa form of the neuraminidase from the non-pathogenic actinomycete *M.viridifaciens* reveals a remarkable similarity in the overall fold of the enzyme to the other neuraminidases characterized so far, in spite of a low sequence identity. However, the technique of multiple isomorphous isomorphous replacement was required to solve the structure, for despite the high structural homology of the 41kDa *M.viridifaciens* neuraminidase with the other neuraminidases, molecular replacement failed to find a solution. This may have been due to the differences in orientation of the β strands.

M.viridifaciens secretes the 41kDa form of the enzyme when colominic acid is used as an inducer, but secretes the 68kDa form when milk casein is used. It is possible that the sialic acid on the casein glycoproteins together with galactose, either on the glycoproteins or the lactose "contaminant" often found with casein, is inducing the larger form of the enzyme. In addition to a 41kDa domain which shows very little difference to the isolated 41kDa enzyme, the 68kDa form of *M.viridifaciens*

neuraminidase possess two extra domains, which are very likely to have been acquired by horizontal gene transfer from a eukaryote. The linker domain appears to have the role of a rigid linker between the other two domains; having a well-packed hydrophobic core and cylindrical shape. A function for the galactose binding domain has yet to be defined, as is the case for the homologous domain in galactose oxidase. However the galactose binding domain may anchor the enzyme to appropriate substrates allowing the catalytic domain to cleave sialic acids. Alternatively the galactose binding domain may work in concert with the catalytic domain, passing appropriate oligosaccharides to the active site through the relative motion of the two domains. This could occur around a flexible pivot point consisting of two adjacent glycine residues (402 and 403) at the C-terminus of the 41kDa domain (Chapter 9).

Two lectin-like domains were previously identified in the neuraminidase from *V.cholerae*. This suggests that this combination of catalytic and carbohydrate binding domains may be common to other members of the neuraminidase family. These additional domains may provide additional targets for ligand design in the case of pathogenic neuraminidases.

In order to further investigate the role of the additional domains in the 68kDa form of the enzyme, several approaches could be taken. Molecular dynamics studies could unveil a route, if indeed possible, by which an oligosaccharide could be transferred to the active site of the catalytic domain. Alternatively, investigating the structure of the enzyme complexed with some non-hydrolysable substrate, such as sialic acid linked to galactose via a Sulphur may reveal the function and mobility of the galactose binding domain of the 68kDa form of *M.viridifaciens* neuraminidase.

Appendix 1

Protein methods

Neuraminidase assay

Neuraminidase activity was assayed using the method of Warren (1959). To each tube 0.1ml of the enzyme sample was added, together with 0.1ml of the substrate fetuin. Fetuin is obtained from fetal calf serum and consists of N-acetyl neuraminic acid attached to an oligosaccharide chain. The solutions were then mixed and placed in a water bath at 37°C for about 30 minutes. After this time, 0.1ml of 0.2M sodium periodate in 9M phosphoric acid was added, the tubes shaken and allowed to stand at room temperature for 20 minutes. 1.0ml of 10% sodium arsenite in a solution of 0.5M sodium sulfate-0.1M sulfuric acid was then added and the tubes shaken until a yellow-brown colour disappeared. 3ml of 0.6% thiobarbituric acid in 0.5M sodium sulfate was added, the tubes shaken, capped with a glass bead and vigorously heated for 15 minutes in a boiling water bath. The tubes were then removed and placed in cold water for 5 minutes. 1ml of each cooled solution was then transferred to another tube which contained 1ml of cyclohexanone. Each tube was shaken twice and then centrifuged for 2-5 minutes at 500-1000r.p.m. Ensuring that the upper butanol layer was free of turbidity, this layer was carefully pipetted into a cuvette and the absorbance read at a wavelength of 549nm in a spectrophotometer. Blanks containing only fetuin, with no added enzyme were used to calibrate the machine.

Whilst the enzyme and substrate are incubating, the neuraminidase is catalysing the cleavage of the terminal sialic acid attached to the oligosaccharide chain of the

fetuin, which produces free N-acetyl neuraminic acid (NANA). The addition of sodium periodate in 9M phosphoric acid provides a strongly acidic environment which is required to obtain a maximal extinction coefficient. Strong acid is also probably required to remove the acetyl or glycolyl from the amino group before oxidation by the periodate. The sodium arsenite is used to destroy the periodate present. The thiobarbituric acid is then coupled with the β-formylpyruvic acid (oxidized N-acetyl neuraminic acid) and the resulting chromophore extracted into cyclohexanone. The presence of sodium sulfate in the sodium arsenite and thiobarbituric acid solutions leads to a 28% increase in the observed molar extinction coefficient by making the extraction of the chromophore by cyclohexanone more complete. The concentration of the chromophore, which is equivalent to the concentration of the sialic acid released by the neuraminidase can then be calculated by the following procedure:

$$A=\varepsilon\, c l \qquad \text{(Beer Lambert Law)}$$

A= absorption of the chromophore measured at 549nm

ε= Molar extinction coefficient of N-acetyl neuraminic acid; =57,000

l=path length (1cm) of cuvette

$$\text{Thus } c = \frac{A}{57}\,\text{mM of NANA produced in 30 mins.}$$

$$= \frac{A}{57 \times 30}\,\text{mM/min}$$

$$= \frac{A}{57x30x\text{tot. protein}} \text{mM/min/mg}$$ where tot.protein is total mg's of protein in the assay.

Protein estimation

The concentration of protein samples was determined spectrophotometrically at 595nm using the method of Bradford (1976). A standard curve was generated for each set of protein estimations, using 0 to 20$\mu g\mu l^{-1}$ solutions of Bovine serum albumin.

SDS-polyacrylamide gel electrophoresis

Discontinuous SDS-polyacrylamide gel electrophoresis was performed according to the method of Laemmli (1970). A concentration of 10% acrylamide was used in the gels.

Gel filtration

(As described in Chapter 2)

Anion exchange

(As described in Chapter 2)

Crystallizations

The method of hanging drop vapour diffusion was employed in all cases (Mcpherson, 1990). Cover slips were silanised before use. All protein samples were concentrated to 10-12mg/ml using the centricon concentrators prior to crystallization

trials. The appropriate volumes of protein sample and precipitant containing buffer were added to each coverslip, which was then inverted and placed over a well in a Limbro tissue culture plate. Each well in use, contained 1ml of the buffer and precipitant solution. An air tight seal between the coverslip and the well was achieved by placing a thin layer of grease around the rim of the well.

Materials

Crude native *M.viridifaciens* neuraminidase supplied as an ammonium sulfate precipitate from a culture filtrate, was a generous gift from Dr. Takayuki Uwajima, Tokyo Research Laboratories, Kyowa Hakko Kogyo Co. Ltd, Tokyo, Japan. Additional crude recombinant *M.viridifaciens* neuraminidase, used to produce additional 41kDa crystals, was a generous gift from Dr. Mamoru Hasegawa, Tokyo Research Laboratories, Kyowa Hakko Kogyo Co. Ltd, Tokyo, Japan.

General reagents were obtained from BDH, SIGMA or FISONS. Neuraminidase from *C.perfringens* (used to verify the validity of the neuraminidase assay) was obtained from Boehringer Manheim.

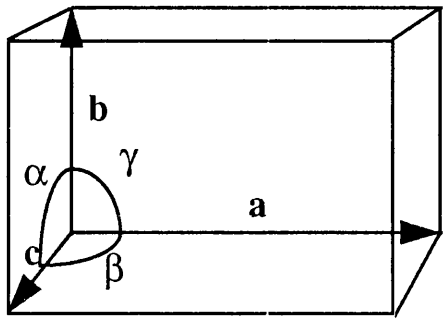
Appendix 2

Diffraction by a set of planes

W.L. Bragg showed that the angles at which diffracted beams emerge from a crystal can be computed by treating diffraction as if it were from sets of equivalent parallel planes of atoms in a crystal. Diffraction occurs from parallel planes present in each unit cell of the crystal.

The unit cell is the smallest and simplest volume element that is completely representative of the whole crystal. The array of points at the vertices of unit cells is known as the lattice. The dimensions of the unit cell are designated by six numbers: the lengths of the three unique edges a, b and c; and three unique angles α, β and γ. There are seven different classes of unit cell, depending on equivalences in the unit cell dimensions. A cell in which $a \neq b \neq c$ and $\alpha \neq \beta \neq \gamma$ is termed triclinic (fig. 2.1). If $a \neq b \neq c$, $\alpha = \gamma = 90°$ and $\beta > 90°$, the cell is monoclinic. If $a = b \neq c$, $\alpha = \beta = 90°$ and $\gamma = 120°$, the cell is hexagonal or trigonal. For cells where $\alpha = \beta = \gamma = 90°$and $a = b = c$, the cell is cubic; if $a = b \neq c$, the cell is tetragonal; and if $a \neq b \neq c$, the cell is orthorhombic.

Fig. 2.1 The Triclinic unit cell.



The 41kDa *M.viridifaciens* neuraminidase crystals are orthorhombic and the 68kDa *M.viridifaciens* neuraminidase crystals are monoclinic.

Regularly spaced planes that can be drawn through lattice points in the crystal are effectively sources of diffraction. The Miller indices *hkl* identify a particular set of equivalent, parallel planes and specify the number of planes that exist per unit cell in the a, b, c directions respectively. The unit cell is in real space, whereas the diffraction pattern is in reciprocal space. The reciprocal lattice, with unit cell lengths a*, b*and c*, is related to the lattice in real space for a unit cell where $\alpha = \beta = \gamma = 90°$ by $a^* = 1/a$, $b^* = 1/b$ and $c^* = 1/c$. The Miller indices are used to designate individual reflections in the reciprocal space of the diffraction pattern.

Bragg`s Law

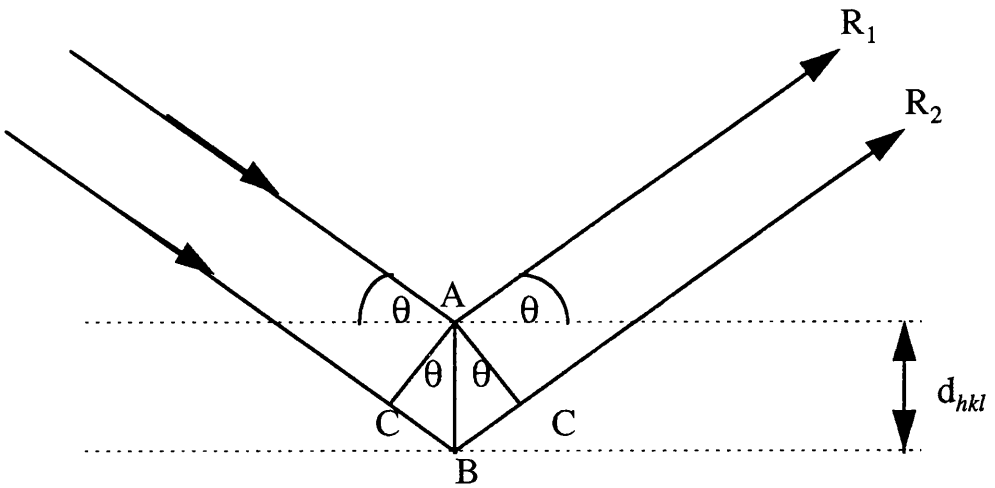
(a) Real space

Bragg demonstrated that a set of parallel planes with Miller indices *hkl* and interplanar spacing d_{hkl} produces a diffracted beam when X-rays of wavelength of λ impinge on the planes at an angle θ and are reflected at the same angle, only if θ meets the condition:

$$n\lambda = 2d_{hkl} \sin\theta \qquad \text{where n is an integer.}$$

Fig. 2.2 illustrates the conditions necessary to produce a diffracted beam. Two rays R_1 and R_2 are reflected from from two parallel planes of lattice points with interplanar spacing d_{hkl} at an angle θ. The additional distance travelled by the more deeply penetrating ray R_2 is $2d_{hkl} \sin\theta$.

Fig. 2.2 Bragg's law in real space. Interplanar spacing = d_{hkl}. Two rays R_1 and R_2 are reflected at an angle θ by two parallel planes of lattice points. $\sin\theta$ = BC/AB. The additional distance travelled by ray R_2 is 2BC and BC = AB $\sin\theta = d_{hkl} \sin\theta$.



(b) Reciprocal space

Bragg's Law can also be applied in reciprocal space. Fig. 2.3 shows an a* b* plane of a reciprocal lattice, with an X-ray beam (X O) impinging on the crystal along this plane. The circle has radius $1/\lambda$, with its centre C on XO and passing through O. $1/\lambda$ is the wavelength of the X-rays in reciprocal space. If the crystal is rotated about O, the reciprocal lattice is also rotated around O , successively bringing reciprocal lattice points such as P and P` into contact with the circle. The angle PBO is θ and hence:

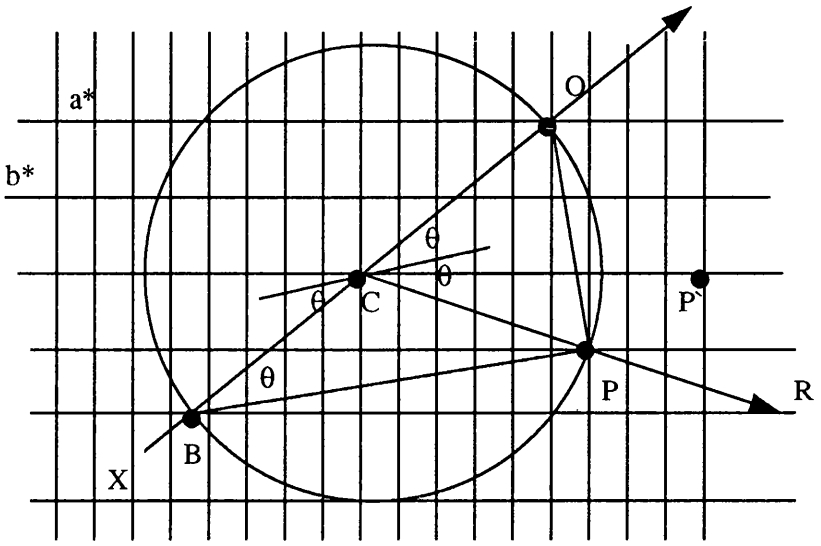
$$\sin\theta = \frac{OP}{BO} = \frac{OP}{2/\lambda}$$

$$\text{and therefore } 2\frac{1}{OP}\sin\theta = \lambda$$

Since P is a reciprocal lattice point, the length of OP is $1/d_{hkl}$ and this gives:

$$n\lambda = 2d_{hkl}\sin\theta$$

Fig. 2.3 Bragg's Law in reciprocal space. When the crystal is rotated so that a reciprocal lattice point such as P comes in contact with the circle of radius $1/\lambda$, Bragg's law is satisfied and ray R emerges from the crystal. Reciprocal lattice point P` does not intersect the circle.



The Fourier series and Fourier transform

A simple wave can be described by a periodic function such as the equations:

$$f(x) = F \cos 2\pi(hx + \alpha)$$

or

$$f(x) = F \sin 2\pi(hx + \alpha)$$

In these functions, f(x) specifies the vertical height of the wave at any horizontal position x along the wave. The variable x and the constant α are angles expressed in fractions of the wavelength. The constant F specifies the amplitude of the wave and h the wavelength or frequency. α is the phase of the wave.

Any wave can be described as the sum of simple waves. This sum is known as a Fourier series and a Fourier series of n terms can be written using the first simple wave equation above as follows:

$$f(x) = F_0 \cos 2\pi(0x + \alpha_0)$$

$$+$$

$$F_1 \cos 2\pi(1x + \alpha_1)$$

$$+$$

$$F_2 \cos 2\pi(2x + \alpha_2)$$

$$+$$

$$\ldots$$

$$+$$

$$F_n \cos 2\pi(nx + \alpha_n)$$

or

$$f(x) = \sum_{h=0}^{n} F_h \cos 2\pi(hx + \alpha h)$$

Using the waveform [cos 2π(hx) + i sin 2π(hx)], the above Fourier series can be re-written thus:

$$f(x) = \sum_{h=0}^{n} F_h[\cos 2\pi(hx) + i \sin 2\pi(hx)]$$

From complex number theory: $\cos\theta + i \sin\theta = e^{i\theta}$, so the Fourier series becomes:

$$f(x) = \sum_{h} F_h \, e^{2\pi i(hx)}$$

This is a Fourier series for a one dimensional wave. In three dimensions this becomes:

$$f(x, y, z) = \sum_{hkl} F_{hkl} \, e^{2\pi i(hx+ky+lz)}$$

Fourier showed that for a function $f(x)$, there is another function $F(h)$ such that:

$$F(h) = \int_{-\infty}^{+\infty} f(x) e^{2\pi i(hx)} dx$$

$F(h)$ is the Fourier transform of $f(x)$ and also $f(x)$ is the transform of $F(h)$. In other words the above Fourier transform operation is reversible:

$$f(x) = \int_{-\infty}^{+\infty} F(h) e^{-2\pi i(hx)} dh$$

Again though we are interested in the Fourier series in three dimensions:

$$F(hkl) = \int_{xyz} f(xyz) e^{2\pi i(hx+ky+lz)} dx\, dy\, dz$$

and $$f(xyz) = \int_{hkl} F(hkl) e^{-2\pi i(hx+ky+lz)} dh\, dk\, dl$$

A structure factor, which describes a given reflection, can also be expressed as a Fourier series. A single term in this series is called an atomic structure factor f_{hkl} and represents the contribution of the single atom j to reflection *hkl*:

$$f_{hkl} = f_j \, e^{2\pi i (hx_j + ky_j + lz_j)}$$

The term f_j is the scattering factor of atom j and is a function that treats the atom as a sphere of electron density. As a result the function is different for each element, since each element has a different number of electrons. x_j, y_j and z_j are the coordinates of atom j in the unit cell.

The structure factor F_{hkl}, for a unit cell containing n atoms, is the sum of diffractive contributions from all atoms in the unit cell, or in other words the sum of all the individual f_{hkl} values. The Fourier series for n terms is given by:

$$F_{hkl} = \sum_{j=1}^{n} f_j e^{2\pi i (hxj + kyj + lzj)}$$

F_{hkl} may also be written as the sum of contributions from each volume element of electron density in the unit cell:

$$F_{hkl} = \int_{hkl} \rho(x\ y\ z) e^{2\pi i (hx + ky + lz)} \, dx\, dy\, dz$$

or

$$F_{hkl} = \int_{V} \rho(x\ y\ z) e^{2\pi i (hx + ky + lz)} \, dV$$

Here, F_{hkl} is the Fourier transform of ρ(x y z) on the set of real lattice planes (*hkl*). *V* is simply the unit cell volume, i.e. the integration is over all values of *x*, *y* and *z* in the unit cell.

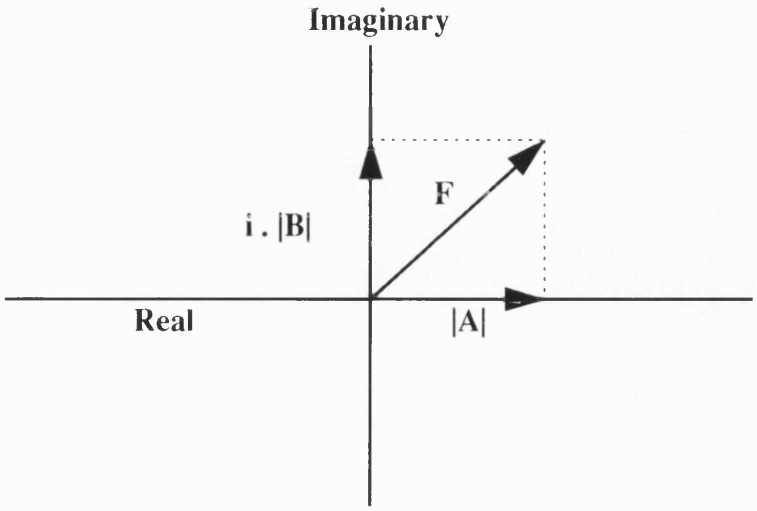
Since the Fourier transform operation is reversible, the electron density can be represented as the transform of the structure factors, thus:

$$\rho\ (x\ y\ z) = \frac{1}{V} \sum_{hkl} F_{hkl}\ e^{-2\pi i(hx+ky+lz)}$$

Here each F_{hkl} is a structure factor, itself a Fourier series, describing a specific reflection in the diffraction pattern. F_{hkl} is a periodic function, possessing amplitude, frequency and phase. The frequency is that of the X-ray source. The amplitude of F_{hkl} is obtained from the diffraction pattern, since the amplitude is proportional to the square root of the intensity of the measured reflection (I_{hkl}). However the phase of F_{hkl} is unknown, since this is lost when the X-rays are diffracted. This is known as the phase problem and is the fundamental problem to be solved in macromolecular crystallography.

Structure factors may be represented as complex vectors (a + ib). The length of this vector represents the amplitude of the structure factor and the phase of the structure factor is represented by the angle α that the vector makes with the real axis, when the origin of the vector is placed at the origin of the complex plane. This is illustrated in fig. 2.4.

Fig. 2.4 The Structure factor **F**, represented in two dimensions by complex numbers.



From fig. 2.4;

$$\cos\alpha = \frac{|\mathbf{A}|}{|\mathbf{F}|} \qquad \text{and} \qquad \sin\alpha = \frac{|\mathbf{B}|}{|\mathbf{F}|}$$

and therefore:

$$|\mathbf{A}| = |\mathbf{F}|\cos\alpha \qquad \text{and} \qquad |\mathbf{B}| = |\mathbf{F}|\sin\alpha$$

Expressing **F** in the form **A** + i**B**;

$$\mathbf{F} = |\mathbf{A}| + i|\mathbf{B}| = |\mathbf{F}|\,(\cos\alpha + i\sin\alpha)$$

Since $\cos\alpha + i\sin\alpha = e^{i\alpha}$

$$\mathbf{F} = |\mathbf{F}|\,e^{i\alpha}$$

At present we have the following electron density equation:

$$\rho(x\ y\ z) = \frac{1}{V}\sum_{hkl} F_{hkl}\, e^{-2\pi i(hx+ky+lz)}$$

If we insert $\mathbf{F} = |\mathbf{F}| e^{i\alpha}$ into the above equation we obtain:

$$\rho (x\ y\ z) = \frac{1}{V} \sum_{hkl} |F_{hkl}|\ e^{i\alpha_{hkl}}\ e^{-2\pi i(hx+ky+lz)}$$

The phase angle in radians is given by: $\alpha = 2\pi\alpha'$. This simplifies the exponential terms to:

$$\rho (x\ y\ z) = \frac{1}{V} \sum_{hkl} |F_{hkl}|\ e^{2\pi i\alpha'_{hkl}}\ e^{-2\pi i(hx+ky+lz)}$$

Finally, if we add the exponents:

$$\rho (x\ y\ z) = \frac{1}{V} \sum_{hkl} |F_{hkl}|\ e^{-2\pi i(hx+ky+lz-\alpha'_{hkl})}$$

This final electron density equation gives the desired electron density as a function of the known amplitudes $|\mathbf{F}|$ and the unknown phases α'_{hkl} of each reflection.

Estimation of the unknown phases

(a) Isomorphous replacement

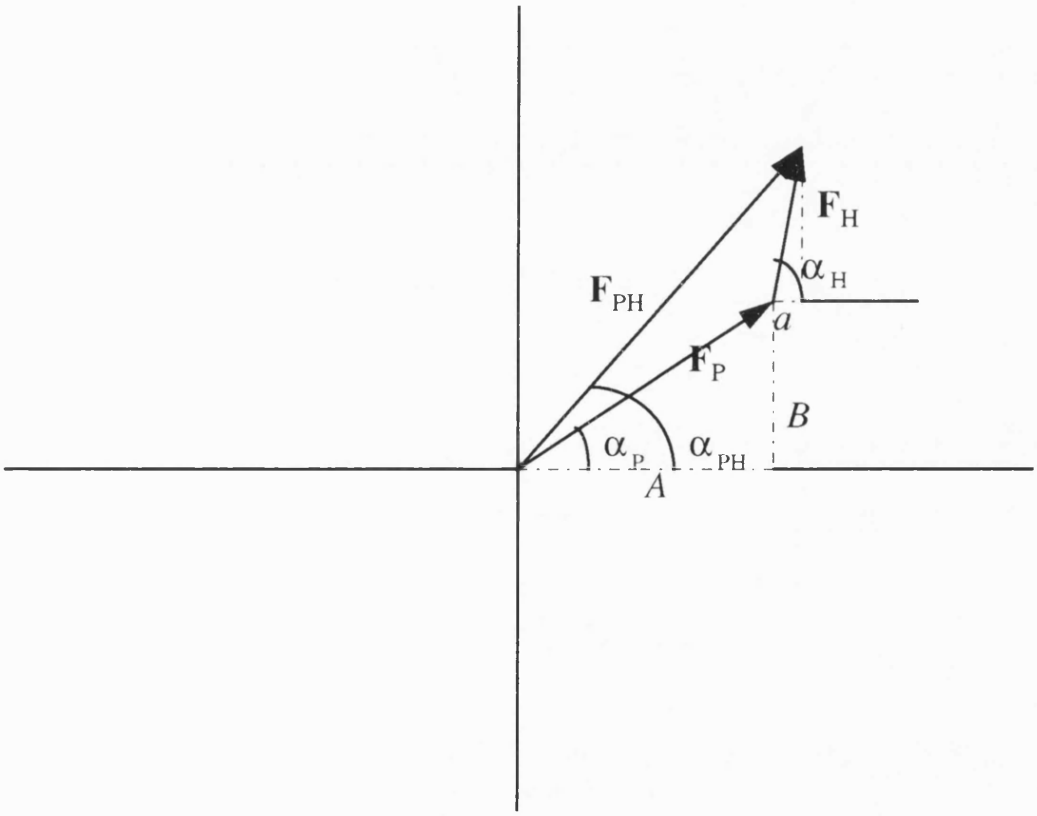
Each atom in the unit cell contributes to every reflection in the diffraction pattern. Hence if one or a very small number of atoms were added to identical sites in all unit cells of a crystal, changes would arise in the diffraction pattern as a result of the additional contributions of the added atom. This slight perturbation in the diffraction pattern caused by the added atom can be used to obtain initial estimates of phases. The added atom must be a strong diffractor, i.e. a heavy atom, in order for these changes in the intensities of the reflections to be measured.

If $\mathbf{F}_P$ is the structure factor of a native reflection, $\mathbf{F}_{PH}$ is the structure factor of the corresponding derivative reflection and $\mathbf{F}_H$ is the structure factor for the heavy atom alone, then because the diffractive contributions of atoms are additive vectors the following relationship holds:

$$\mathbf{F}_{PH} = \mathbf{F}_H + \mathbf{F}_P$$

Fig. 2.5 illustrates the above relationship in vector form:

Fig. 2.5 The phase triangle: the structure factor $\mathbf{F}_{PH}$ is the sum of contributions from the native ($\mathbf{F}_P$) and heavy atom ($\mathbf{F}_H$) structure factors.



The position of the heavy atoms in the unit cell may be calculated using a Fourier series called the Patterson function P(u,v,w). The Patterson synthesis can be constructed from intensity measurements, even though no phase information is available. The amplitude of each term in the Patterson function is the square of one structure factor, which is proportional to the measured reflection intensity. The difference Patterson function for the heavy atoms alone has amplitudes $\Delta F^2 = (|F_{HP}| - |F_P|)^2$. Thus the difference Patterson function is:

$$\Delta P\,(u\ v\ w) = \frac{1}{V} \sum_{hkl} \Delta F^2_{h\,k\,l}\ \ e^{-2\pi i(hu + kv + lw)}$$

From the location of the heavy atoms, the vector F_H can be calculated and therefore the phase of the native protein alone (α_P) can be solved. In the vector diagram of fig. 2.5, α_P can be determined from the cosine law:

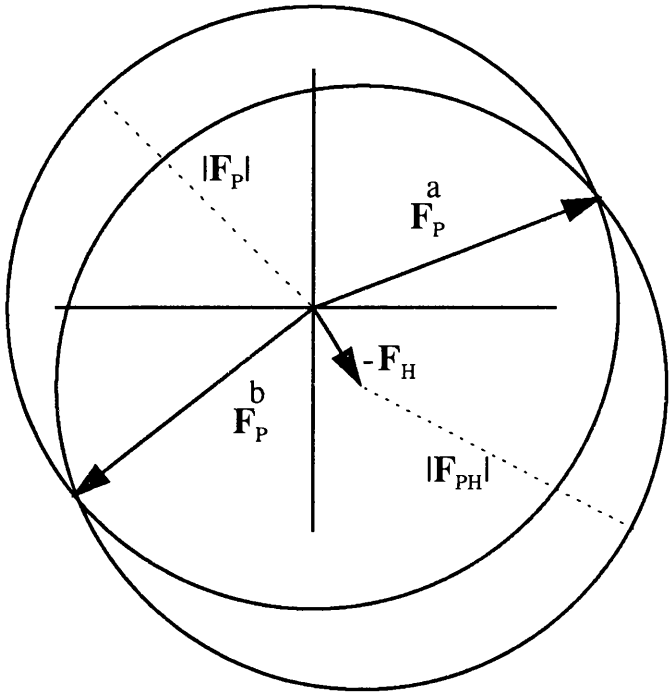
$$\alpha_P = \alpha_H + \cos^{-1}\left(\frac{F^2_{PH} - F^2_P - F^2_H}{2F_P F_H}\right) = \alpha_H \pm \alpha'$$

As can be seen from the above calculation of α_P, there are two possible values for α_P which cannot be distinguished with one isomorphous derivative. In addition the initial phase estimates are inaccurate due to errors which arise as a result in measurement of amplitudes for example. These errors result in lack of closure, where the phase triangle does not close. In order to improve the qualtiy of the phase estimates and to resolve the phase ambiguity which arises with one derivative, at least one further derivative is usually required.

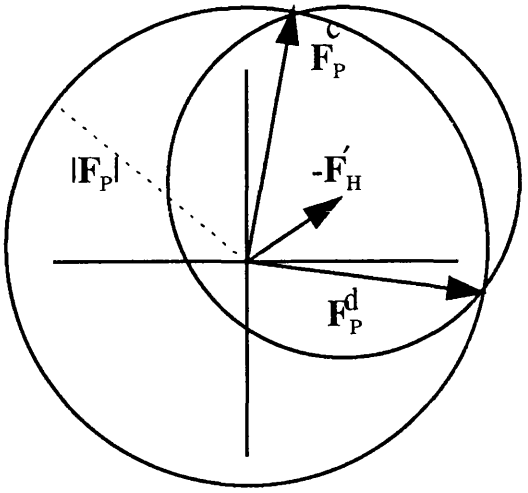
The Harker construction shown in fig. 2.6 best explains how this phase ambiguity is overcome. In fig. 2.6 (a) the vector $-\mathbf{F}_H$ is placed at the origin and a circle of radius $|\mathbf{F}_{HP}|$ is drawn, centred on the head of vector $-\mathbf{F}_H$. Next a circle of radius $|\mathbf{F}_P|$ is added, centred at the origin. The equation $\mathbf{F}_P = \mathbf{F}_{PH} - \mathbf{F}_H$ is only true at points where the two circles intersect. Thus the correct phase angle α_P is either that for the vector $\mathbf{F}_p^a$ or $\mathbf{F}_p^b$. Fig. 2.6 (b) shows the phase determination using a second heavy-atom derivative. $\mathbf{F}_H'$ is the structure factor for the second heavy atom. The radius of the smaller circle is $|\mathbf{F}'_{HP}|$, the amplitude of $\mathbf{F}'_{HP}$ for the second heavy atom derivative. For this derivative, $\mathbf{F}_P = \mathbf{F}'_{HP} - \mathbf{F}'_H$. In fig. 2.6 (c), the phase circles from fig's 2.6 (a) and (b) are superimposed, showing that the correct phase angle is that for $\mathbf{F}_P^C$ and $\mathbf{F}_P^a$ which are identical.

Fig. 2.6 Harker circle constructions **(a)** Vectors from the origin to intersections of the two intersections are solutions to $\mathbf{F}_P = \mathbf{F}_{PH} - \mathbf{F}_H$ **(b)** A second heavy atom derivative indicates two possible phases; $\mathbf{F}_P^d$ or $\mathbf{F}_P^C$ **(c)** Superposition of the previous phase circles indicates the correct phase for this reflection.

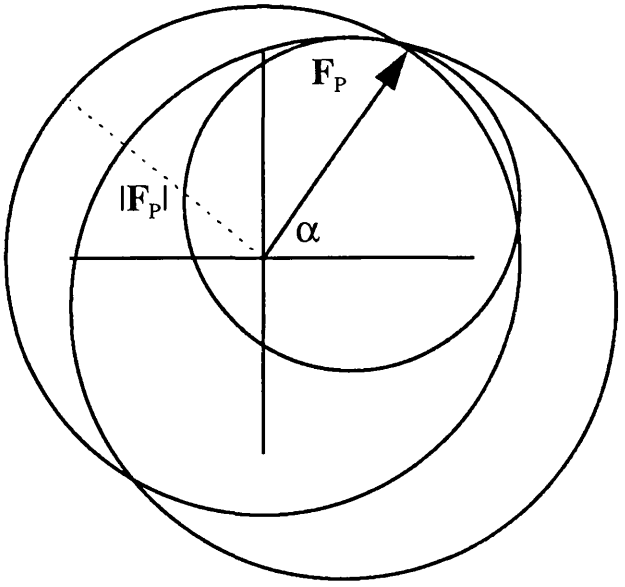
(a)



(b)


c
$\mathbf{F}_P$
$|\mathbf{F}_P|$
$-\mathbf{F}'_H$
$\mathbf{F}_P^d$

(c)


$\mathbf{F}_P$
$|\mathbf{F}_P|$
α

Various statistical parameters are used to judge the quality of the phase estimates, such as figure of merit, lack of closure errors and various R-factors, and to weight the Fourier syntheses towards the more accurately determined phases, whilst culling or down-weighting those phases that are more questionable. The best electron density map obtained therefore, is that with centroid phases (PHIBEST) and weighted by FOM (figure of merit - Chapter 4). Due to errors in measurements of intensities, lack of isomorphism and incorrect heavy atom positions, the three phase circles in fig. 2.6 (c) will not intersect in one place in practice and the phase determination will be ambiguous. Crick and Blow have given a thorough treatment of errors in the method of isomorphous replacement (Blow & Crick, 1959) and best choice of weights to use in the calculation of the electron density map.

Anomalous scattering also provides a means of obtaining phase information. This is where Friedel's Law breaks down and the reflections hkl and -h-k-l are no longer equal in intensity. Anomalous scattering arises when the added heavy atom's absorption edge is near the X-ray wavelength.

(b) Molecular replacement

The initial determination of phases by the technique of molecular replacement is often attempted if the structure of a similar or homologous macromolecule is known. This homologous structure is known as the search model and is placed in the unit cell of the target crystal and rotated and translated in order to obtain the best agreement between the observed and calculated diffraction data. The rotation and translation

searches for the best orientation and position of the search model are generally performed separately, due to limitations in computing power.

A second molecular replacement strategy used to gain phase information is the search for the orientation of non-crystallographic symmetry elements in the asymmetric unit of the crystal. Using this powerful technique, electron density maps have been improved by "averaging" around local symmetry elements. The former molecular replacement strategy will be discussed in more detail here.

The Rotation function

The rotation function as developed by Rossmann and Blow (1962) defined a Patterson overlap function as the product of two Pattersons integrated over a volume containing only self-vectors, where:

$$R[\Omega] = \int_U P_o(u) \, . \, P_c(u') \, du$$

In the above equation:

u' = $[\Omega]\, u$

Ω = rotation matrix

$P_o(u)$ = value of observed Patterson function at point u

$P_c(u')$ = value of calculated Patterson function at point u'

U volume over which matching is expected

P (u) can be expanded in terms of structure factor amplitudes:

$$P(u) = \frac{1}{V}\sum_{hkl} |F(hkl)|^2 \exp(-2\pi i(hu + kv + lw))$$

As a result, this simplifies R[Ω] to:

$$R[\Omega] = \frac{u}{v^3}\sum_{h}\sum_{p} |F(h)|^2 |F(p)|^2 G_{hh'}$$

In this equation:

$|F(h)|$ = observed structure factor amplitudes

$|F(p)|$ = calculated structure factor amplitudes

h' = $[\Omega]\, p$

$G_{hh'}$ = interference function. This contains information about the rotation in lattice vector h'

This overlap function would be expected to give a high correlation when the Pattersons of the search model and unknown molecule are superimposed.

Crowther (1972) has expanded the Pattersons in the rotation function in terms of spherical harmonics. By representing the Patterson density within a spherical volume as a function of normalized spherical Bessel functions and normalized spherical harmonics, rather than Cartesian Fourier components $|F(h)|^2$, it was possible to use fast Fourier transforms to calculate values of the Rossmann and Blow function. This

resulted in a decrease in the calculation time, but at the expense of limiting the volume of integration to being spherical.

Navaza (1987) implemented the enhancement of rotation peaks in the AMoRE package, by skipping low angular resolution contributions. The formulation of the rotation function is essentially that of Crowther, but uses numerical integration instead of Fourier-Bessel expansions in the radial variable. Important numerical instabilities were removed by the use of a new algorithm to calculate the reduced rotation matrices (Navaza, 1990).

The Translation function

Once the search model has been placed in the correct orientation by a correct solution to the rotation function, the translational parameters are obtained for positioning the model inside the unit cell, using the translation function. The T2 function of Crowther and Blow (1967) is a Patterson overlap function. This function calculates a three dimensional Patterson overlap function using the cross-vectors (intermolecular vectors) contained in the unit cell. The T2 function has the form:

$$T2\,(t) = \int_V (P_{o(u)} - P_{m(u)}) \,.\, (P_{c(u,\,t)} - P_{m(u)})\, du$$

In this equation:

$P_{o\,(u)}$ = Patterson of unknown structure from $|F_o|$'s

$P_{m(u)}$ = sum of separate self-Pattersons of n asymmetric units of the known structure in their correct orientations, obtained from model |F|'s.

$P_{c(u,t)}$ = Patterson of the complete unit cell of the trial structure consisting of x asymmetric units of known structure in their correct orientation and position.

V = Unit cell volume

In order to obtain a large correlation from the overlap function, the translational components of the search model (t) are varied. The overlap function can also be expressed in terms of structure factors in reciprocal space:

$$T(2) = \sum_{h} \left(F_o^2(h) - F_m^2(h)\right) \cdot \left(F_c^2(h,t) - F_m^2(h)\right)$$

The output from the T(2) translation function is a list of modified Fourier coefficients which are then used to calculate a three dimensional translation function map using the fast Fourier transform (Ten Eyck, 1973).

Refinement

Refinement is the process of altering the parameters of the model in order to improve the agreement between the model and the experimental observations. These parameters are in the most part the positions (x_j, y_j, z_j) and the thermal parameters (B_j) of the j atoms. A function which describes the discrepancy between the amplitudes of

the observed reflections and the values calculated from the model parameters is the function to be minimized. There are a variety of refinement packages available, implementing different methods of refinement, including X-PLOR (Brunger, 1988) and TNT (Tronrud *et al.*, 1987).

The X-PLOR energy function

In general, the energy function to be minimized in the refinement process can be represented by the equation:

$$E_{total} = E_{empirical} + E_{x\text{-}ray}$$

where $E_{x\text{-}ray} = \Sigma\, w\, (|F_o| - |F_c|)^2$, the X-ray pseudo energy term.

$$\text{and } E_{empirical} = \sum_{bonds} k(r - r_0)^2 \quad (1)$$

$$+ \sum_{angles} (k_\theta(\theta - \theta_0)^2 + k_{ub}(r_{13} - r_{ub})^2) \quad (2)$$

$$+ \sum_{dihedrals,impropers\ i=1,\,m} \begin{cases} k_\phi(1 + \cos(n\phi + \sigma)) \text{ if } n > 0 \\ k_\phi(\phi - \sigma)^2 \text{ if } n = 0 \end{cases} \quad (3)$$

$$+ \sum_{i<j} \left(\frac{A_i}{r_{AD}} - \frac{B_j}{r_{AD}}\right) \quad (4)$$

OR $\cos^m(\theta_{A\text{-}H\text{-}D}), \cos^n(\theta_{AA\text{-}A\text{-}H})\ SW(r^2_{AD}, r^2h_{on}, r^2h_{off})$

OR $SW[\cos^2(\theta_{A\text{-}D\text{-}H}), \cos^2(\theta_{hon}), \cos^2(\theta_{hoff})]$

$$+ \quad \frac{A}{R^{12}} - \frac{B}{R^{6}} = 4\xi\left(\left(\frac{\sigma}{R}\right)^{12} - \left(\frac{\sigma}{R}\right)^{6}\right)H\left(R - R_{CUT}\right) \qquad (5)$$

OR $\left(\frac{A}{R^{12}} - \frac{B}{R^{6}}\right)SW(R, R_{ON}, R_{OFF})$

OR $C_{rep}\,(\max(0, (k^{rep}R_{min})^{irexp} - R^{irexp}))^{rexp}$

$$+ \quad {}^{Q_i Q_j}\frac{c}{\xi_0 R}\,\text{heavy}(R - Rcut)$$

OR $${}^{Q_i Q_j}\frac{c}{\xi_0 R}\left(1 - \frac{R^2}{R_{off}^2}\right)^2 \qquad (6)$$

OR $${}^{Q_i Q_j}\frac{c}{\xi_0 R}\,SW\,(R, R_{on}, R_{off})$$

OR 0

$$+ \quad \sum_{i<j} f_{ELEC}\,(R_{ij}) + e_{14} \sum_{(i,j)\in\{1\text{-}4\}} f_{ELEC}\,(R_{ij}) \qquad (7)$$

$$+ \quad \sum_{i<j} f_{VDW}\,(R_{ij}) + \sum_{(i,j)\in\{1\text{-}4\}} f_{VDW}\,(R_{ij})$$

$$+ \sum_{s=1}^{ns} \sum_{i<j} f_{VDW}(R_i s_j) \qquad (8)$$

$$+ \sum_{s=1}^{ns} \sum_{i<j} f_{ELEC}(R_i s_j)$$

$$+ \sum_{S \in NCS} \sum_{i<j} f_{VDW}(r_i - Sr_j) \qquad (9)$$

$$+ \sum_{S \in NCS} \sum_{i<j} f_{ELEC}(r_i - Sr_j)$$

The term (**1**), describes the covalent bond energy, where r is the actual bond length, k is an energy constant and r_o is equilibrium constants. Term (**2**) describes the bond angle energy, where k_θ and k_{ub} are energy constants and θ_0 and r_{ub} are equilibrium constants. θ is the value of the angle and r_{13} is the distance between the first and third atom defining the angle. The term (**3**) describes energy terms involving dihedral and improper angles. ϕ is the torsion angle, k_ϕ are energy constants, n are periodicities, m are multiplicities and δ are the phase shifts. The explicit hydrogen-bond energy term, $E_{H\ BON}$ is given by the term in (**4**). The atoms AA, A, H and D are the receptor antecedent, acceptor, hydrogen and donor heavy atom respectively, where i, j and m are positive integers and SW is the switching function, which has the form:

$$SW(R, R_{on}, R_{off}) = \begin{cases} 0 & \text{if } R > Roff \\ \dfrac{(R^2 - R_{off}^2 * (R^2 - R_{off}^2 - 3(R^2 - R_{on}^2)))}{R_{off}^2 - R_{on}^2} & \text{if } R_{off} > R > R_{on} \\ 1 & \text{if } R < Ron \end{cases}$$

The switching function SW defined above, maintains a continuous and smooth function to the cutoff point so that derivatives are well defined at the cutoff point. The van der Waals function is given by term (**5**). The first equation provides the truncation option, the second the switched option and the third the repel option. H is the heavy-side function and SW is the same switching function as described above. The van der Waals function is described by a Lennard-Jones potential in both the truncated and switched option. Term (6) gives the electrostatic function. The first equation is for pure truncation, the second gives the shifted option, the third provides for the 1/R option and 0 gives the repel option. These options are used to truncate the electrostatic potential; because of the need to limit the number of pair interactions and to avoid discontinuities in the forces. Intramolecular interactions are provided for by term (7). The summation is over the individual nonbonded interaction energies for pairs of atoms (i, j) within the current molecular structure. In order to reduce computational time, there is a cutoff, which is the distance R_{ij} at which the energy becomes zero, given by R_{off}, for both switched and shifted nonbonded options. Term (8) computes the crystallographic symmetry interactions between the molecule(s) located in the asymmetric unit and all symmetry-related molecules (i.e. intermolecular interactions). The first equation computes symmetry interactions for packing van der Waals forces and the second for packing electrostatic energies. The final term (9) computes the

interaction between non-crystallographically related molecules, using the f_{VDW} and f_{ELEC} functional forms.

Stages in the X-PLOR refinement

Fig. 2.7 summarizes the following stages in the X-PLOR refinement. The first stage in the X-PLOR refinement was GENERATE. GENERATE adds the nitrogen hydrogens and oxygen hydrogens to the model, prior to a round of refinement and prepares a molecular structure file, which contains information such as bond angles, atom names and types and hydrogen bonding terms. CHECK determines the ideal weight W_A between E_{XREF} and $E_{EMPIRICAL}$ and also checks the initial R value for the structure. The weight is determined by running a brief molecular dynamics calculation in the absence of E_{XREF} and then comparing the norm of the gradient of $E_{EMPIRICAL}$ and the norm of the gradient of E_{XREF}. PREPSTAGE uses the weight calculated in CHECK and performs several cycles of minimization on the model in order to relieve strain or bad contacts. PREPSTAGE is a typical least squares refinement, where the function to be minimized is $\sum w\ (|F_{obs}|-|F_{calc}|)^2$; simultaneously with restrictive stereochemical terms. However, as with all least squares only the nearest local minimum from the starting coordinates will be reached. Therefore PREPSTAGE was followed by refinement in SLOWCOOL. SLOWCOOL, or simulated annealing involves refinement using the principles of molecular dynamics. Molecular dynamics consists of simultaneously solving Newton's equations of motion for all atoms under the influence of a potential derived from empirical stereochemical terms and observational restraints. In SLOWCOOL, the initial temperature of the system is set to typically 2000K, in order to apply kinetic energy to the system. Random velocities, whose distribution is a

function of the temperature, are applied to all the atoms. This results in a new position and velocity for each atom a very small time step later. The high kinetic energy of the system, derived from the high initial temperature, allows the overcoming of energy barriers and a large number of conformations can be explored. In the SLOWCOOL algorithm, the temperature of the system is gradually reduced from 2000K in the hope of finding the global energy minimum.

SLOWCOOL ouputs a refined set of coordinates and its R-factor. The R-factor is defined by the equation:

$$R = \frac{\sum ||F_{OBS}| - |F_{CALC}||}{\sum |F_{OBS}|}$$

The R-factor is a common method of monitoring the progress of a refinement. The better the agreement between the structure factor amplitudes from the observed reflections and those calculated from the model, the lower the R-factor will be and presumably the better the model. R-values approaching 12% have been achieved for the most extensively refined models with resolution better than 2Å and R-values in the vicinity of 60% indicate a structure that does not fit the data any better than does a random one.

Despite the stereochemical restraints imposed in the refinement process it is still possible to overfit the diffraction data, particularly if the observation to parameter ratio is poor. Thus an incorrect model can be refined to a fairly good R value as past examples have shown (Branden & Jones, 1990). An unbiased test of the accuracy of an atomic model was therefore proposed. A statistical quantity R_{free} was defined that

measures the agreement between the observed and calculated structure factor amplitudes for a "test" set of reflections that is omitted in the refinement and modelling process (Brunger, 1992). Examples have shown that there is a high correlation between the accuracy of the phases calculated for the atomic model and the R_{free} (Brunger, 1992). The unique set of observed reflections is separated into a "test" set *T* and a "working" set *A* and the free R value to be computed is:

$$R_{free} = \frac{\sum_{(hkl) \in T} ||Fobs(hkl)| - k|Fcalc(hkl)||}{\sum_{(hkl) \in T} |Fobs(hkl)|}$$

T is omitted in the modelling and refinement process and the crystallographic residual to be minimized is:

$$R'_A = \sum_{(hkl) \in A} (|Fobs(hkl)| - k|Fcalc(hkl)|)^2$$

Temperature factor refinement

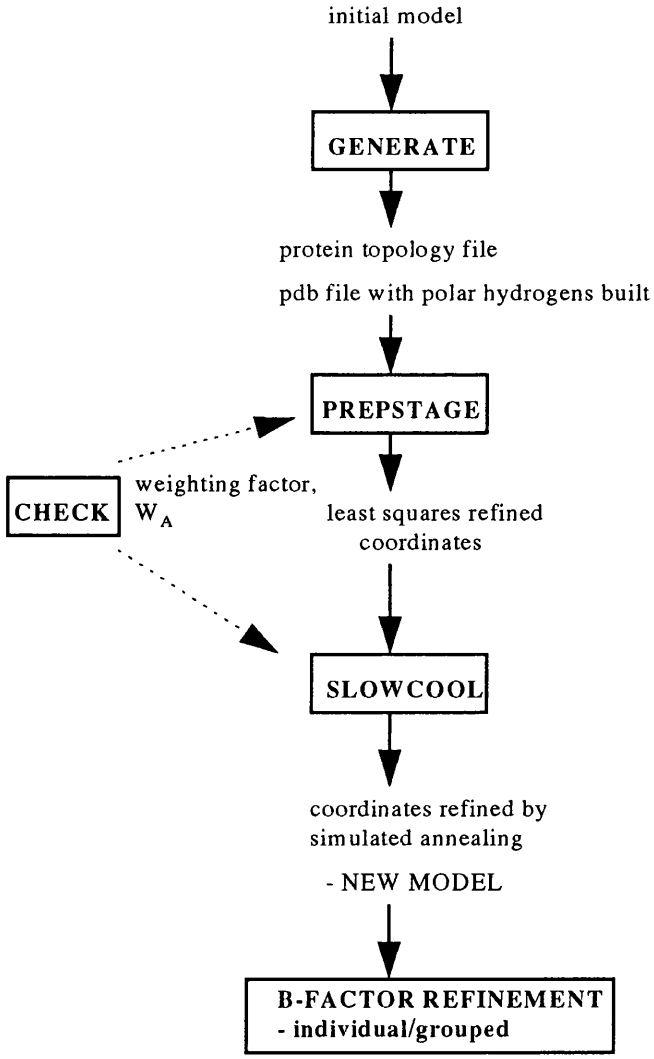
Assuming that the occupancies for the set of coordinates are correct, B_j is a measure of the thermal motion at atom *j*. In the simplest case of purely harmonic thermal motion of equal magnitude in all directions, or isotropic vibration, B_j has the following relationship to magnitude of vibration:

$$B_j = 8\pi^2 \{u_j^2\} = 79\{u_j^2\}$$

where $\{u_j^2\}$ is the mean-square displacement of the atom from its rest position in $Å^2$.

In order to refine the temperature factors (B-factors) there is an essential requirement for a high observation to parameter ratio; i.e. a high completeness of data at a high resolution. Isotropic B-factors can be refined at a resolution of 2.7Å or better, with a good completeness of data. However, in order to refine B-factors anisotropically requires six parameters per atom and therefore a very high resolution with a high completeness of data is required.

Fig. 2.7 Summary of the refinement stages employed in X-PLOR.



References

Abergel, C., Nesa, M.P., Fontecilla - Camps, J.C. (1991) *J. Cryst. Growth* **110**, 11 - 19.

Aisaka, K. & Uwajima, T. (1987) *FEMS Microbiol. Lett.* **44**, 289 - 291.

Aisaka, K., Igarashi, A., Uwajima, T. (1991) *Agric. Biol. Chem.* **55**, 997 - 1004.

Ashwell, G. & Morell, A. (1974) *Adv. Enzymol.* **41**, 99 - 128.

Bork, P. & Doolittle, R.F. (1994) *J. Mol. Biol.* **236**, 1277 - 1282.

Bork, P., Holm, L., Sander, C. (1994) *J. Mol. Biol.* **242**, 309 - 320.

Bouwstra, J.B., Deyl, C.M., Vliegenthart, J.F.G. (1987) *Biol. Chem. Hoppe - Seyler* **368**, 269.

Bradford, M.M. (1976) *Anal. Biochem.* **72**, 248 - 254.

Branden, C.I. & Jones, A. (1990) *Nature* **343**, 687 - 689.

Brunger, A.T. (1992) *Nature* **355**, 472 - 475.

Brunger, A.T. (1988) *J. Mol. Biol.* **203**, 803 - 816.

Brunger, A.T. (1988) X-PLOR manual (version 5.9).

Burmeister, W.P., Ruigrok, R.W.H., Cusack, S. (1992) *EMBO* **11**, 49 - 56.

Cacalano, G., Kays, M., Saiman, L., Prince, A. (1992) *J. Clin. Invest.* **89**, 1866 - 1874.

Cassidy, J.T., Jourdian, G.W., Roseman, S. (1965) *J. Biol. Chem.* **240**, 3501.

Castellano, E.E., Oliva, G., Navaza, J. (1992) *J. Appl. Cryst.* **25**, 281 - 284.

CCP4 (1979). The SERC (U.K.) Collaborative computing project No. 4: A suite of programs for protein crystallography (Daresbury Lab., Warrington, U.K.).

Chong, A.K.J., Pegg, M.S., Taylor, N.R., von Itzstein, M., (1992) *Eur. J. Biochem.* **207**, 335 - 343.

Choppin, P.W. & Scheid, A. (1980) *Rev. Infect. Dis.* **2**, 40 - 61.

Colman, P.M., Hoyne, P.A., Lawrence, M.C. (1993) *J. Virol.* **67**, 2972 - 2980.

Corfield, A.P., Wember, M., Schauer, R., Rott, R. (1982) *Eur. J. Biochem.* **124**, 521 - 525.

Corfield, A.P., Sander - Wewer, M., Veh, R.W., Wember, M., Schauer, R. (1986) *Biol. Chem. Hoppe - Seyler* **367**, 433 - 439.

Corfield, A.P., Wagner, S.A., Clamp, J.R., Kriaris, M.S., Hoskins, L.C. (1992) *Infection and Immunity* **60**, 3971 - 3978.

Corfield, A.P., Wagner, S.A., O'Donnell, L.J.D., Durdey, P., Mountford, R.A., Clamp, J.R. (1993) *Glycoconj. J.* **10**, 72 - 81.

Cowtan, K.D. & Main, P. (1993) *Acta. Cryst.* **D49** 148 - 157.

Crennell, S.J., Garman, E.F., Laver, W.G., Vimr, E.R., Taylor, G.L. (1993) *Proc. Natl. Acad. Sci.* **90,** 9852 - 9856.

Crennell, S., Garman, E., Laver, G., Vimr, E., Taylor, G. (1994) *Structure* **2**, 535 - 544.

Crowther, R.A. & Blow, D.M. (1967) *Acta. Cryst.* **23**, 544 - 548.

Crowther, R.A. (1972) In *The Molecular Replacement Method* Ed. Rossmann, M.G. New York, 174 - 178.

Cudney, B., Patel, S., McPherson, A. (1994) *Acta. Cryst.* **D50**, 479 - 483.

Doolittle, R.F. & Bork, P. (1993) Sci. Amer. **269**, 50 - 56.

Driscoll, P.C., Cyster, J.G., Campbell, I.D., Williams, A.F. (1991) *Nature* **353**, 762 - 765.

Drzeniek, R. (1972) *Curr. Top. Microbiol.* **59**, 35.

Ferrari, J., Harris, R., Warner, T.G. (1994) *Glycobiology* **4**, 367 - 373.

Flashner, M., Kessler., J., Tanenbaum, S.W. (1983) *Archs. Biochem. Biophys.* **221,** 188 - 196.

Galen, J.E., Ketley, J.M., Fasano, A., Richardson, S.H., Wasserman, S.S., Kaper, J.B. (1992) *Infect. Immun.* **60**, 406 - 415.

GCG suite of programs. *Program manual for the Wisconsin package, Version 8, September 1994, Genetics Computer Group, 575 Science Drive, Madison, Wisconsin, USA 53711.*

Glanzer, B.I., Gyorgydeak, Z., Bernet, B., Vasella, A. (1991) *Helv. chim. Acta.* **74,** 343 - 369.

Gottschalk, A. & Lind, P.E. (1949) *Nature* **164**, 232 - 233.

Gottschalk. A. (1966) In *The Glycoproteins. Their Composition, Structure and Function.*

Guo, X. & Sinnott, M.L. (1993) *Biochem. J.* **296**, 291 - 292.

Guo, X. & Sinnott, M.L. (1993) *Biochem. J.* **294**, 653 - 656.

Hasegawa, M., Dairi, T., Ohta, T., Hashimoto, E. (1991) *J. Bacteriol.* **173**, 7004 - 7011.

Hendrickson, W.A. (1991) *Science* **254**, 51 - 58.

Hirst, G.K. (1941) *Science* **94**, 22 - 23.

Holmgren, A. & Branden, C-I. (1989) *Nature* **342**, 248 - 251.

Hoskins, L.C., Agustines, M., McKee, W.B., Boulding, E.T., Kriaris, M., Niedermeyer, G. (1985) *J. Clin Invest.* **75**, 944 - 953.

Hoyer, L.L., Roggentin, P., Schauer, R., Vimr, E.R. (1991) *J. Biochem.* **110**, 462 - 467.

Hoyer, L.L., Hamilton, A.C., Steenbergen, S.M., Vimr, E.R. (1992) *Mol. Microbiol.* **6**, 873 - 884.

Hsu, M.C., Scheid, A., Choppin, P.W. (1981) *J. Biol. Chem.* **256**, 3557 - 3563.

Iorio, R.M. and Glickman, R.L. (1992) *J. Virol.* **66**, 6626 - 6633.

Ito, N., Phillips, S.E.V., Stevens, C., Zumrut, B., Ogel, Z.B., McPherson, M.J., Keen, J.N., Yadav, K.D.S., Knowles, P.F. (1991) *Nature* **350**, 87 - 90.

Ito, N., Phillips, S.E.V., Yadav, K.D.S., Knowles, P.F. (1994) *J. Mol. Biol.* **238**, 794 - 814.

Janakiraman, M.N., White, C.L., Laver, W.G., Air, G.M., Luo, M. (1994) *Biochemistry* **33**, 8172-8179.

Jancarik, J. & Kim, S-H. (1991) *J. Appl. Cryst.* **24**, 409 - 411.

Jazin, E.E., Luquetti, A.O., Rassi, A., Frasch, A.C.C. (1991) *Infect. Immun.* **59**, 2188 - 2191.

Jones, T.A., Zou, J.Y., Cowan, S.W., Kjeldgaard, M. (1991) *Acta. Cryst.* **A47**, 110 - 119.

Jones, E.Y., Davis, S.J., Williams, A.F., Harlos, K., Stuart, D.I. (1992) *Nature* **360**, 232 - 239.

Kabsch, W. & Sander, C. (1983) *Biopolymers* **22**, 2577 - 2637.

Kabsch, W. (1988) *J. Appl. Cryst.* **21**, 916 - 924.

Kingsbury, D.W., Bratt, M.A., Choppin, P.W., Hanson, R.P., Hosaka, Y., ter Meulen, V., Norrby, E., Plowright, W., Rott, R., Wunner, W.H. (1978) *Intervirology* **10**, 137 - 152.

Kottgen, E., Bauer, C., Reutter, W., Gerok, W. (1979) *Klin. Wochenschr.* **57**, 151 - 159, 199 - 214.

Kraut, J., Sieker, L.C., High, D.F., Freer, S.T. (1962) *Proc. Nat. Acad. Sci. USA* **48**, 1417 - 1424.

Laemmli, U.K. (1970) *Nature* **227**, 680 - 685.

Landolfi, N.F., Leone, J., Womack, J.E., Cook, R.G. (1985) *Immunogenetics* **22**, 159 - 167.

Leahy, D.J., Axel, R., Hendrickson, W.A. (1992a) *Cell* **68**, 1145 - 1162.

Leahy, D.J., Hendrickson, W.A., Aukhil, I., Erickson, H.P. (1992b) *Science* **258**, 987 - 991.

Little, E., Bork, P., Doolittle, R.F. (1994) *J. Mol. Evol.* **39**, 631 - 643.

Luthy, R., Bowie, J.U., Eisenberg, D. (1992) *Nature* **356**, 83 - 85.

Luzatti, V. (1952) *Acta. Cryst.* **5**, 802.

Maehr, H. & Schaffner, C.P. (1967) *J. Chromatogr.* **30**, 572 - 578.

Markwell, M.A. & Fox, C.F. (1980) *J.Virol.* **33**, 152 - 166.

Matthews, B.W. (1968) *J. Mol. Biol.* **33**, 491 - 497.

McPherson, A. (1982) In *The preparation and analysis of protein crystals* Ed. Wiley, J. and Sons, New York.

McPherson, A. (1989) *Scientific American* **260**, 62 - 69.

Meindl, P. & Tuppy, H. (1969) *Hoppe - Seyler's Z. physiol. Chem.* **350**, 1088-1092.

Meindl, P., Bodo, G., Palese, P., Schulman, J., Tuppy, H. (1974) *Virology* **58**, 457-463.

Milligan, T.W., Mattingly, S.J., Straus, D.C. (1980) *J. Bacteriol.* **144**, 164 - 172.

Ming, M., Chuenkova, M., Ortega-Barria, E., Pereira, M.E.A. (1993) *Mol. Biochem. Parasitol.* **59**, 243 - 252.

Miyagi, T. & Tsuiki, S. (1984) *Eur. J. Biochem.* **141**, 75 - 81.

Miyagi, T. & Tsuiki, S. (1985) *J. Biol. Chem.* **260**, 6710 - 6716.

Miyagi, T., Sagawa, J., Konno, K., Tsuiki, S. (1990) *J. Biol. Chem.* **107**, 794 - 798.

Morris, A.L., MacArthur, M.W., Thornton, J.N. (1992) *Proteins* **12**, 345 - 364.

Morrison, T. and Portner, A. (1991) In *The Paramyxoviruses* Ed. Kingsbury, D.W. Plenum Publishing Co., New York, 347 - 382.

Muller, H.E. (1976) *Zentrabl. Bakteriol. Parasitenkd. Infektionskr. Hyg. Abt. 1 Reihe A* **235**, 106 - 110.

Murphy, B.R. & Webster, R.G. (1990) In *Virology* Ed. Fields, B.N., Knipe, D.M. *et al.* Raven Press Ltd., New York, **Chapter 40**, 1091-1152.

Musser, J.M., Mattingly, S.J., Quentin, R., Goudeau, A., Selander, R.K. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 4731 - 4735.

Nara, T., Yamamoto, M., Kawamoto, I., Takayama, K, Okachi, R., Takasawa, S., Sato, S., Sato, T. (1977) *J. Antibiot.* **30**, 533 - 540.

Navaza, J. (1987) *Acta. Cryst.* **A43**, 645 - 653.

Navaza, J. (1992) AMoRE In *Molecular replacement* Ed. Dodson, E.J., Glover, S. Wolf, W. SERC Daresbury Laboratory, Warrington, UK. 87 - 90.

Nees, S., Veh, R.W., Schauer, R. (1975) *Hoppe-Seyler's Z. Physiol. Chem.* **356,** 1027 - 1042.

Nojiri, H., Takaku, F., Tetsuka, T., Saito, M. (1982) *Biochem. Biophys. Res. Commun.* **104**, 1239 - 1246.

Okachi, R., Kawamoto, I., Takasawa, S., Yamamoto, M., Sato, S., Sato, T., Nara, T. (1974) *J. Antibiot.* **27**, 793 - 800.

Otwinowski, Z. (1993) In *Proceedings of the CCP4 study weekend* Ed. Sawyer, L., Issacs, N., Bailey, S. SERC, Daresbury Laboratory, UK, 56 - 62.

Paulson, J.C., Rogers, G.N., Carroll, S.M., Higa, H.H., Pritchett, T., Milks, G., Sabesan, S. (1984) *Pure & Appl. Chem.* **56**, 797 - 805.

Pai, E.F. (1992) *Curr. Opin. Struct. Biol.* **2**, 821 - 827.

Palese, P., Tobita, K., Ueda, M., Compans, R.W. (1974) *Virology* **61**, 397 - 410.

Pereira, M.E.A., Mejia, J.S., Ortega-Barria, E., Matzilevich, D., Prioli, R.P. (1991) *J. Exp. Med.* **174**, 179 - 191.

Ramachandran, G.N. & Sasisekharan, V. (1968) *Adv. Prot. Chem.* **23**, 283.

Reutter, W., Kottgen, E., Bauer, C., Gerok, W. (1982) In *Sialic acids*: *Chemistry, Metabolism and Function (Cell Biology Monograph)* Ed. Schauer, R. **10,** 263 - 305.

Richardson, J.S. (1981) *Advan. Protein. Chem.* **34**. 167 - 339.

Roelcke, D. (1984) *Funkt. Biol. Med.* **3**, 106 - 126.

Rosenberg, I.A., Prioli, R.P., Ortega-Barria, E., Pereira, M.E.A. (1991) *Mol. Biochem. Parasitol.* **46**, 303 - 306.

Rossmann, M.G., Moras, D., Olsen, K.W. (1974) *Nature* **250**, 194 - 199.

Sagawa, J., Miyagi, T., Tsuiki, S. (1990) *J. Biochem.* **107**, 452 - 456.

Saito, M. & Yu, R.K. (1992) *J. Neurochem.* **58**, 83 - 87.

Sakurada, K., Ohta, T., Hasegawa, M. (1992) *J. Bacteriol.* **174**, 6896 - 6903.

Schauer, R. (1982) *Adv. Carbohydr. Chem. Biochem* **40**, 131 - 234.

Schauer, R. (1983) *Biochem. Soc. Transac.* **11**, 270-271.

Schauer, R., Shukla, A.K., Schroder, C., Muller, E. (1984) *Pure & Appl. Chem.* **56**, 907 - 921.

Scheid, A. & Choppin, P.W. (1974) *Virology* **57**, 475 - 490.

Schenkman, S., Pontes de Carvalho, L., Nussenzweig, V. (1992) *J. Exp. Med.* **175**, 567 - 575.

Smith, M.W., Feng, D-F., Doolittle, R.F. (1992) *TIBS* **17**, 489 - 493.

Sodhi, A. & Prasab, S.B. (1985) *Experientia* **41**, 93 -95.

Spataro, A.C., Morgan, H.R., Bosmann, H.B. (1975) *Proc. Soc. Exp. Biol. Med.* **149**, 486 - 492.

Sutcliffe, M.J., Haneef, I., Carney, D., Blundell, T.L. (1987) *Protein Eng.* **1**, 377 - 384.

Suzuki, J., Murakami, K., Nishimura, Y. (1993) *J. Carbohydr. Chem.* **12**, 201 - 208.

Taira, S. and Nariuchi, H. (1988) *J. Immunol.* **141**, 440 - 446.

Taylor, G., Dineley, L., Glowka, M., Laver, G. (1992) *J. Mol. Biol.* **225**, 1135 - 1136.

Taylor, N.R. & von Itzstein, M. (1994) *J. Med. Chem.* **37**, 616 - 624.

Thompson, S.D., Laver, W.G., Murti, K.G., Portner, A. (1988) *J.Virol.* **62**, 4653 - 4660.

Tronrud, D.E., Ten Eyck, L.F., Matthews, B.W. (1987) *Acta. Cryst.* **A43**, 489 - 501.

Tulip, W.R., Varghese, J.N., Baker, A.T., van Donkelaar, A., Laver, W.G., Webster R.G., Colman, P.M. (1991) *J. Mol. Biol.* **221**, 487 - 497.

Tulsiani, D.R.P. & Carubelli, R. (1970) *J. Biol Chem.* **245**, 1821 - 1827.

Uskui, S., Hoops, P., Sweeley, C.C. (1988a) *J. Biol. Chem.* **263**, 595 - 610.

Uskui, S., Lyu, S., Sweeley, C.C. (1988b) *J. Biol. Chem.* **263**, 6847 - 6853.

Vandekerckhove, F., Schenkman, S., de Carvalho, L.P., Tomlinson, S., Kiso, M., Yoshida, M., Hasegawa, A., Nussenzweig, V. (1992) *Glycobiology* **2**, 541 - 548.

Varghese, J.N., Laver, W.G., Colman, P.M. (1983) *Nature* **303**, 35 - 40.

Venerando, B., Cestaro, B., Fiorilli, A., Ghidoni, R., Preti, A., Tettamanti, G. (1982) *Biochem. J.* **203**, 735 - 742.

Vimr, E.R. & Hoyer, L.L. (1992) *Conference Phillipe Laudat 1991* INSERM.

Wang, B-C. (1985) In *Methods of Enzymology*. Academic Press Inc. **115**, 90 - 113.

Warren, L. (1959) *J. Biol. Chem* **234**, 1971 - 1975.

Weinstein, M.J., Marquez, J.A., Testa, R.T., Wagman, G.H., Oden, E.M., Waitz, J.A. (1970) *J. Antibiot.* **23**, 551 - 554.

Whiteheart, S. W., McLenithan, J.C., Hart, G.W. (1990) *Cell Immunol.* **125**, 337 - 353.

Wilson, J.C., Angus, D.I., von Itzstein, M. (1995) *J. Am. Chem. Soc.* **117**, 4214 - 4217.

Zeigler, M., Sury, V., Bach, G. (1989) *Eur. J. Biochem.* **183**, 455 - 458.

References

Abergel, C., Nesa, M.P., Fontecilla - Camps, J.C. (1991) *J. Cryst. Growth* **110**, 11 - 19.

Aisaka, K. & Uwajima, T. (1987) *FEMS Microbiol. Lett.* **44**, 289 - 291.

Aisaka, K., Igarashi, A., Uwajima, T. (1991) *Agric. Biol. Chem.* **55**, 997 - 1004.

Ashwell, G. & Morell, A. (1974) *Adv. Enzymol.* **41**, 99 - 128.

Bork, P. & Doolittle, R.F. (1994) *J. Mol. Biol.* **236**, 1277 - 1282.

Bork, P., Holm, L., Sander, C. (1994) *J. Mol. Biol.* **242**, 309 - 320.

Bouwstra, J.B., Deyl, C.M., Vliegenthart, J.F.G. (1987) *Biol. Chem. Hoppe - Seyler* **368**, 269.

Bradford, M.M. (1976) *Anal. Biochem.* **72**, 248 - 254.

Branden, C.I. & Jones, A. (1990) *Nature* **343**, 687 - 689.

Brunger, A.T. (1992) *Nature* **355**, 472 - 475.

Brunger, A.T. (1988) *J. Mol. Biol.* **203**, 803 - 816.

Brunger, A.T. (1988) X-PLOR manual (version 5.9).

Burmeister, W.P., Ruigrok, R.W.H., Cusack, S. (1992) *EMBO* **11**, 49 - 56.

Cacalano, G., Kays, M., Saiman, L., Prince, A. (1992) *J. Clin. Invest.* **89**, 1866 - 1874.

Cassidy, J.T., Jourdian, G.W., Roseman, S. (1965) *J. Biol. Chem.* **240**, 3501.

Castellano, E.E., Oliva, G., Navaza, J. (1992) *J. Appl. Cryst.* **25**, 281 - 284.

CCP4 (1979). The SERC (U.K.) Collaborative computing project No. 4: A suite of programs for protein crystallography (Daresbury Lab., Warrington, U.K.).

Chong, A.K.J., Pegg, M.S., Taylor, N.R., von Itzstein, M., (1992) *Eur. J. Biochem.* **207**, 335 - 343.

Choppin, P.W. & Scheid, A. (1980) *Rev. Infect. Dis.* **2**, 40 - 61.

Colman, P.M., Hoyne, P.A., Lawrence, M.C. (1993) *J. Virol.* **67**, 2972 - 2980.

Corfield, A.P., Wember, M., Schauer, R., Rott, R. (1982) *Eur. J. Biochem.* **124**, 521 - 525.

Corfield, A.P., Sander - Wewer, M., Veh, R.W., Wember, M., Schauer, R. (1986) *Biol. Chem. Hoppe - Seyler* **367**, 433 - 439.

Corfield, A.P., Wagner, S.A., Clamp, J.R., Kriaris, M.S., Hoskins, L.C. (1992) *Infection and Immunity* **60**, 3971 - 3978.

Corfield, A.P., Wagner, S.A., O'Donnell, L.J.D., Durdey, P., Mountford, R.A., Clamp, J.R. (1993) *Glycoconj. J.* **10**, 72 - 81.

Cowtan, K.D. & Main, P. (1993) *Acta. Cryst.* **D49** 148 - 157.

Crennell, S.J., Garman, E.F., Laver, W.G., Vimr, E.R., Taylor, G.L. (1993) *Proc. Natl. Acad. Sci.* **90,** 9852 - 9856.

Crennell, S., Garman, E., Laver, G., Vimr, E., Taylor, G. (1994) *Structure* **2**, 535 - 544.

Crowther, R.A. & Blow, D.M. (1967) *Acta. Cryst.* **23**, 544 - 548.

Crowther, R.A. (1972) In *The Molecular Replacement Method* Ed. Rossmann, M.G. New York, 174 - 178.

Cudney, B., Patel, S., McPherson, A. (1994) *Acta. Cryst.* **D50**, 479 - 483.

Doolittle, R.F. & Bork, P. (1993) Sci. Amer. **269**, 50 - 56.

Driscoll, P.C., Cyster, J.G., Campbell, I.D., Williams, A.F. (1991) *Nature* **353**, 762 - 765.

Drzeniek, R. (1972) *Curr. Top. Microbiol.* **59**, 35.

Ferrari, J., Harris, R., Warner, T.G. (1994) *Glycobiology* **4**, 367 - 373.

Flashner, M., Kessler., J., Tanenbaum, S.W. (1983) *Archs. Biochem. Biophys.* **221**, 188 - 196.

Galen, J.E., Ketley, J.M., Fasano, A., Richardson, S.H., Wasserman, S.S., Kaper, J.B. (1992) *Infect. Immun.* **60**, 406 - 415.

GCG suite of programs. *Program manual for the Wisconsin package, Version 8, September 1994, Genetics Computer Group, 575 Science Drive, Madison, Wisconsin, USA 53711.*

Glanzer, B.I., Gyorgydeak, Z., Bernet, B., Vasella, A. (1991) *Helv. chim. Acta.* **74**, 343 - 369.

Gottschalk, A. & Lind, P.E. (1949) *Nature* **164**, 232 - 233.

Gottschalk. A. (1966) In *The Glycoproteins. Their Composition, Structure and Function.*

Guo, X. & Sinnott, M.L. (1993) *Biochem. J.* **296**, 291 - 292.

Guo, X. & Sinnott, M.L. (1993) *Biochem. J.* **294**, 653 - 656.

Hasegawa, M., Dairi, T., Ohta, T., Hashimoto, E. (1991) *J. Bacteriol.* **173**, 7004 - 7011.

Hendrickson, W.A. (1991) *Science* **254**, 51 - 58.

Hirst, G.K. (1941) *Science* **94**, 22 - 23.

Holmgren, A. & Branden, C-I. (1989) *Nature* **342**, 248 - 251.

Hoskins, L.C., Agustines, M., McKee, W.B., Boulding, E.T., Kriaris, M., Niedermeyer, G. (1985) *J. Clin Invest.* **75**, 944 - 953.

Hoyer, L.L., Roggentin, P., Schauer, R., Vimr, E.R. (1991) *J. Biochem.* **110**, 462 - 467.

Hoyer, L.L., Hamilton, A.C., Steenbergen, S.M., Vimr, E.R. (1992) *Mol. Microbiol.* **6**, 873 - 884.

Hsu, M.C., Scheid, A., Choppin, P.W. (1981) *J. Biol. Chem.* **256**, 3557 - 3563.

Iorio, R.M. and Glickman, R.L. (1992) *J. Virol.* **66**, 6626 - 6633.

Ito, N., Phillips, S.E.V., Stevens, C., Zumrut, B., Ogel, Z.B., McPherson, M.J., Keen, J.N., Yadav, K.D.S., Knowles, P.F. (1991) *Nature* **350**, 87 - 90.

Ito, N., Phillips, S.E.V., Yadav, K.D.S., Knowles, P.F. (1994) *J. Mol. Biol.* **238**, 794 - 814.

Janakiraman, M.N., White, C.L., Laver, W.G., Air, G.M., Luo, M. (1994) *Biochemistry* **33**, 8172-8179.

Jancarik, J. & Kim, S-H. (1991) *J. Appl. Cryst.* **24**, 409 - 411.

Jazin, E.E., Luquetti, A.O., Rassi, A., Frasch, A.C.C. (1991) *Infect. Immun.* **59**, 2188 - 2191.

Jones, T.A., Zou, J.Y., Cowan, S.W., Kjeldgaard, M. (1991) *Acta. Cryst.* **A47**, 110 - 119.

Jones, E.Y., Davis, S.J., Williams, A.F., Harlos, K., Stuart, D.I. (1992) *Nature* **360**, 232 - 239.

Kabsch, W. & Sander, C. (1983) *Biopolymers* **22**, 2577 - 2637.

Kabsch, W. (1988) *J. Appl. Cryst.* **21**, 916 - 924.

Kingsbury, D.W., Bratt, M.A., Choppin, P.W., Hanson, R.P., Hosaka, Y., ter Meulen, V., Norrby, E., Plowright, W., Rott, R., Wunner, W.H. (1978) *Intervirology* **10**, 137 - 152.

Kottgen, E., Bauer, C., Reutter, W., Gerok, W. (1979) *Klin. Wochenschr.* **57**, 151 - 159, 199 - 214.

Kraut, J., Sieker, L.C., High, D.F., Freer, S.T. (1962) *Proc. Nat. Acad. Sci. USA* **48**, 1417 - 1424.

Laemmli, U.K. (1970) *Nature* **227**, 680 - 685.

Landolfi, N.F., Leone, J., Womack, J.E., Cook, R.G. (1985) *Immunogenetics* **22**, 159 - 167.

Leahy, D.J., Axel, R., Hendrickson, W.A. (1992a) *Cell* **68**, 1145 - 1162.

Leahy, D.J., Hendrickson, W.A., Aukhil, I., Erickson, H.P. (1992b) *Science* **258**, 987 - 991.

Little, E., Bork, P., Doolittle, R.F. (1994) *J. Mol. Evol.* **39**, 631 - 643.

Luthy, R., Bowie, J.U., Eisenberg, D. (1992) *Nature* **356**, 83 - 85.

Luzatti, V. (1952) *Acta. Cryst.* **5**, 802.

Maehr, H. & Schaffner, C.P. (1967) *J. Chromatogr.* **30**, 572 - 578.

Markwell, M.A. & Fox, C.F. (1980) *J.Virol.* **33**, 152 - 166.

Matthews, B.W. (1968) *J. Mol. Biol.* **33**, 491 - 497.

McPherson, A. (1982) In *The preparation and analysis of protein crystals* Ed. Wiley, J. and Sons, New York.

McPherson, A. (1989) *Scientific American* **260**, 62 - 69.

Meindl, P. & Tuppy, H. (1969) *Hoppe - Seyler's Z. physiol. Chem.* **350**, 1088-1092.

Meindl, P., Bodo, G., Palese, P., Schulman, J., Tuppy, H. (1974) *Virology* **58**, 457-463.

Milligan, T.W., Mattingly, S.J., Straus, D.C. (1980) *J. Bacteriol.* **144**, 164 - 172.

Ming, M., Chuenkova, M., Ortega-Barria, E., Pereira, M.E.A. (1993) *Mol. Biochem. Parasitol.* **59**, 243 - 252.

Miyagi, T. & Tsuiki, S. (1984) *Eur. J. Biochem.* **141**, 75 - 81.

Miyagi, T. & Tsuiki, S. (1985) *J. Biol. Chem.* **260**, 6710 - 6716.

Miyagi, T., Sagawa, J., Konno, K., Tsuiki, S. (1990) *J. Biol. Chem.* **107**, 794 - 798.

Morris, A.L., MacArthur, M.W., Thornton, J.N. (1992) *Proteins* **12**, 345 - 364.

Morrison, T. and Portner, A. (1991) In *The Paramyxoviruses* Ed. Kingsbury, D.W. Plenum Publishing Co., New York, 347 - 382.

Muller, H.E. (1976) *Zentrabl. Bakteriol. Parasitenkd. Infektionskr. Hyg. Abt. 1 Reihe A* **235**, 106 - 110.

Murphy, B.R. & Webster, R.G. (1990) In *Virology* Ed. Fields, B.N., Knipe, D.M. *et al.* Raven Press Ltd., New York, **Chapter 40**, 1091-1152.

Musser, J.M., Mattingly, S.J., Quentin, R., Goudeau, A., Selander, R.K. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 4731 - 4735.

Nara, T., Yamamoto, M., Kawamoto, I., Takayama, K, Okachi, R., Takasawa, S., Sato, S., Sato, T. (1977) *J. Antibiot.* **30**, 533 - 540.

Navaza, J. (1987) *Acta. Cryst.* **A43**, 645 - 653.

Navaza, J. (1992) AMoRE In *Molecular replacement* Ed. Dodson, E.J., Glover, S. Wolf, W. SERC Daresbury Laboratory, Warrington, UK. 87 - 90.

Nees, S., Veh, R.W., Schauer, R. (1975) *Hoppe-Seyler's Z. Physiol. Chem.* **356,** 1027 - 1042.

Nojiri, H., Takaku, F., Tetsuka, T., Saito, M. (1982) *Biochem. Biophys. Res. Commun.* **104**, 1239 - 1246.

Okachi, R., Kawamoto, I., Takasawa, S., Yamamoto, M., Sato, S., Sato, T., Nara, T. (1974) *J. Antibiot.* **27**, 793 - 800.

Otwinowski, Z. (1993) In *Proceedings of the CCP4 study weekend* Ed. Sawyer, L., Issacs, N., Bailey, S. SERC, Daresbury Laboratory, UK, 56 - 62.

Paulson, J.C., Rogers, G.N., Carroll, S.M., Higa, H.H., Pritchett, T., Milks, G., Sabesan, S. (1984) *Pure & Appl. Chem.* **56**, 797 - 805.

Pai, E.F. (1992) *Curr. Opin. Struct. Biol.* **2**, 821 - 827.

Palese, P., Tobita, K., Ueda, M., Compans, R.W. (1974) *Virology* **61**, 397 - 410.

Pereira, M.E.A., Mejia, J.S., Ortega-Barria, E., Matzilevich, D., Prioli, R.P. (1991) *J. Exp. Med.* **174**, 179 - 191.

Ramachandran, G.N. & Sasisekharan, V. (1968) *Adv. Prot. Chem.* **23**, 283.

Reutter, W., Kottgen, E., Bauer, C., Gerok, W. (1982) In *Sialic acids*: *Chemistry, Metabolism and Function (Cell Biology Monograph)* Ed. Schauer, R. **10,** 263 - 305.

Richardson, J.S. (1981) *Advan. Protein. Chem.* **34**. 167 - 339.

Roelcke, D. (1984) *Funkt. Biol. Med.* **3**, 106 - 126.

Rosenberg, I.A., Prioli, R.P., Ortega-Barria, E., Pereira, M.E.A. (1991) *Mol. Biochem. Parasitol.* **46**, 303 - 306.

Rossmann, M.G., Moras, D., Olsen, K.W. (1974) *Nature* **250**, 194 - 199.

Sagawa, J., Miyagi, T., Tsuiki, S. (1990) *J. Biochem.* **107**, 452 - 456.

Saito, M. & Yu, R.K. (1992) *J. Neurochem.* **58**, 83 - 87.

Sakurada, K., Ohta, T., Hasegawa, M. (1992) *J. Bacteriol.* **174**, 6896 - 6903.

Schauer, R. (1982) *Adv. Carbohydr. Chem. Biochem* **40**, 131 - 234.

Schauer, R. (1983) *Biochem. Soc. Transac.* **11**, 270-271.

Schauer, R., Shukla, A.K., Schroder, C., Muller, E. (1984) *Pure & Appl. Chem.* **56**, 907 - 921.

Scheid, A. & Choppin, P.W. (1974) *Virology* **57**, 475 - 490.

Schenkman, S., Pontes de Carvalho, L., Nussenzweig, V. (1992) *J. Exp. Med.* **175**, 567 - 575.

Smith, M.W., Feng, D-F., Doolittle, R.F. (1992) *TIBS* **17**, 489 - 493.

Sodhi, A. & Prasab, S.B. (1985) *Experientia* **41**, 93 -95.

Spataro, A.C., Morgan, H.R., Bosmann, H.B. (1975) *Proc. Soc. Exp. Biol. Med.* **149**, 486 - 492.

Sutcliffe, M.J., Haneef, I., Carney, D., Blundell, T.L. (1987) *Protein Eng.* **1**, 377 - 384.

Suzuki, J., Murakami, K., Nishimura, Y. (1993) *J. Carbohydr. Chem.* **12**, 201 - 208.

Taira, S. and Nariuchi, H. (1988) *J. Immunol.* **141**, 440 - 446.

Taylor, G., Dineley, L., Glowka, M., Laver, G. (1992) *J. Mol. Biol.* **225**, 1135 - 1136.

Taylor, N.R. & von Itzstein, M. (1994) *J. Med. Chem.* **37**, 616 - 624.

Thompson, S.D., Laver, W.G., Murti, K.G., Portner, A. (1988) *J.Virol.* **62**, 4653 - 4660.

Tronrud, D.E., Ten Eyck, L.F., Matthews, B.W. (1987) *Acta. Cryst.* **A43**, 489 - 501.

Tulip, W.R., Varghese, J.N., Baker, A.T., van Donkelaar, A., Laver, W.G., Webster R.G., Colman, P.M. (1991) *J. Mol. Biol.* **221**, 487 - 497.

Tulsiani, D.R.P. & Carubelli, R. (1970) *J. Biol Chem.* **245**, 1821 - 1827.

Uskui, S., Hoops, P., Sweeley, C.C. (1988a) *J. Biol. Chem.* **263**, 595 - 610.

Uskui, S., Lyu, S., Sweeley, C.C. (1988b) *J. Biol. Chem.* **263**, 6847 - 6853.

Vandekerckhove, F., Schenkman, S., de Carvalho, L.P., Tomlinson, S., Kiso, M., Yoshida, M., Hasegawa, A., Nussenzweig, V. (1992) *Glycobiology* **2**, 541 - 548.

Varghese, J.N., Laver, W.G., Colman, P.M. (1983) *Nature* **303**, 35 - 40.

Venerando, B., Cestaro, B., Fiorilli, A., Ghidoni, R., Preti, A., Tettamanti, G. (1982) *Biochem. J.* **203**, 735 - 742.

Vimr, E.R. & Hoyer, L.L. (1992) *Conference Phillipe Laudat 1991* INSERM.

Wang, B-C. (1985) In *Methods of Enzymology*. Academic Press Inc. **115**, 90 - 113.

Warren, L. (1959) *J. Biol. Chem* **234**, 1971 - 1975.

Weinstein, M.J., Marquez, J.A., Testa, R.T., Wagman, G.H., Oden, E.M., Waitz, J.A. (1970) *J. Antibiot.* **23**, 551 - 554.

Whiteheart, S. W., McLenithan, J.C., Hart, G.W. (1990) *Cell Immunol.* **125**, 337 - 353.

Wilson, J.C., Angus, D.I., von Itzstein, M. (1995) *J. Am. Chem. Soc.* **117**, 4214 - 4217.

Zeigler, M., Sury, V., Bach, G. (1989) *Eur. J. Biochem.* **183**, 455 - 458.